Universidad Politécnica de Madrid

Escuela Técnica Superior
de Ingenieros de Telecomunicación

Segmentación de secuencias de vídeo: análisis y desarrollo de estrategias para la detección de cambios de toma y para la detección de objetos móviles

Tesis Doctoral

Carlos Cuevas Rodríguez
Ingeniero de Telecomunicación
2011

Departamento de Señales, Sistemas y Radiocomunicaciones

Escuela Técnica Superior de Ingenieros de Telecomunicación

Segmentación de secuencias de vídeo: análisis y desarrollo de estrategias para la detección de cambios de toma y para la detección de objetos móviles

Tesis Doctoral

Autor:
Carlos Cuevas Rodríguez
Ingeniero de Telecomunicación
Universidad Politécnica de Madrid

Director:
Narciso García Santos
Doctor Ingeniero de Telecomunicación
Catedrático del Dpto. de Señales, Sistemas y Radiocomunicaciones
Universidad Politécnica de Madrid

2011

TESIS DOCTORAL

Segmentación de secuencias de vídeo: análisis y desarrollo de estrategias para la detección de cambios de toma y para la detección de objetos móviles

Autor: Carlos Cuevas Rodríguez
Director: Narciso García Santos

Tribunal nombrado por el Mgfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día . . . . de . . . . . de 2011.

Presidente: D. . . . . .

Vocal: D. . . . . .

Vocal: D. . . . . .

Vocal: D. . . . . .

Secretario: D. . . . . .

Realizado el acto de defensa y lectura de la Tesis el día . . . . de . . . . . de 2011 en . . . . . .

Calificación: . . . . . .

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

A Beatriz

Agradecimientos

Son muchas las personas que me han ayudado a llegar hasta aquí y por eso quiero dedicar estos párrafos para agradecérselo.

En primer lugar quiero dar las gracias a mi mujer, Beatriz, por ser la que me ha dado las fuerzas para seguir adelante en los momentos más difíciles. Junto a ella, mis padres, Pedro y Conchi, mi hermana Marta y mi cuñado Pedro Pablo han sido los que más han tenido que aguantar mis excentricidades y pesimismos y, a pesar de eso, siempre me han animado, apoyado y querido. Por eso también les doy las gracias especialmente a ellos. Tampoco puedo olvidarme de mis abuelos, tíos y tías, ya que si he llegado hasta aquí es también gracias a su apoyo y confianza en mí.

También quiero dar las gracias a la gente de mi grupo (el GTI). En primer lugar a mi director de tesis, Narciso, por todo el apoyo que me ha dado y todo lo que me ha enseñado. Asimismo he de acordarme de Luis, ya que con él fue con quien empecé mi trayectoria como investigador; y de Fernando, del que siempre he obtenido ayuda para resolver cualquier problema que me haya podido surgir. Tampoco puedo pasar por alto el agradecimiento a todos mis compañeros en el GTI, especialmente a aquellos que han estado conmigo desde que comencé la tesis: gracias Marcos, Carlos Roberto, Raúl, Víctor y Jon, gracias a todos por no limitaros a ser buenos compañeros y por haberos convertido en grandes amigos, gracias por haberme ofrecido siempre vuestra ayuda y por haberme hecho pasar momentos inolvidables a lo largo de estos años. Para terminar quiero aclarar que, aunque no los he nombrado específicamente por ser demasiados, no olvido al resto de compañeros que me han acompañado durante la realización de la tesis, tanto los que ya dejaron el GTI como los que continúan en él.

Gracias a todos.

Resumen

En los últimos años se han producido importantes avances tecnológicos que han dado lugar a la demanda de nuevas aplicaciones relacionadas con el análisis de secuencias de vídeo. Por un lado ha aumentado el número de aplicaciones de edición de vídeo que requieren estrategias capaces de segmentar los vídeos en tomas. Por otro lado, los dispositivos de última generación que utilizan cámaras requieren aplicaciones que hacen uso técnicas de detección de objetos móviles. En esta tesis se proponen distintas estrategias de segmentación en tomas y de detección de objetos móviles que cumplen los requisitos de calidad, velocidad y facilidad de uso demandados por los usuarios.

En primer lugar se propone un sistema rápido y eficaz, capaz segmentar las secuencias de vídeo en tomas. Para detectar las transiciones abruptas entre tomas se lleva a cabo un análisis de diferencias entre imágenes consecutivas. En paralelo, para detectar las transiciones graduales se analizan los bordes a lo largo de las imágenes. En último lugar, gracias a una estrategia basada en el análisis del movimiento entre pares de imágenes, se localizan y se descartan la mayor parte de las falsas detecciones.

Por otro lado, se propone una estrategia de detección, basada en el popular método de mezcla de gaussianas, capaz de adaptar dinámicamente el número de gaussianas utilizadas en cada instante, en función de las variaciones recientes sufridas por cada píxel. La estrategia propuesta mejora muy notablemente la eficiencia computacional de otros métodos similares y, además, reduce la dependencia de los resultados con los parámetros del método original.

Como alternativa a esta estrategia, para mejorar los resultados en escenarios en los que el fondo sufre grandes variaciones, se proponen otros esquemas de detección basados en el modelado no paramétrico del fondo y del primer plano. Para obtener detecciones de calidad, independientemente de las características de la secuencia analizada, el ancho de las funciones locales utilizadas en dichos modelados se estima dinámicamente. Además, gracias a una estrategia de seguimiento basada en un filtro de partículas, se obtienen importantes mejoras computacionales y de calidad. Adicionalmente, utilizando una innovadora combinación de color normalizado y gradientes se consigue reducir el número de falsas detecciones debidas a sombras y reflejos.

Tras la aplicación de las estrategias propuestas sobre distintas bases de datos compuestas por secuencias con contenido crítico, tanto para la detección de transiciones como para la detección de objetos móviles, se ha comprobado que son capaces de ofrecer resultados de gran calidad y que, además, mejoran la eficiencia computacional de otras estrategias similares.

Abstract

The important technological advances experienced along the last years have resulted in an important demand for new and efficient computer vision applications. On the one hand, the increasing use of video editing software has given rise to a necessity for faster and more efficient editing tools that, in a first step, perform a temporal segmentation in shots. On the other hand, the number of electronic devices with integrated cameras has grown enormously. These devices require new, fast, and efficient computer vision applications that include moving object detection strategies.

In this dissertation, we propose a temporal segmentation strategy and several moving object detection strategies, which are suitable for the last generation of computer vision applications requiring both low computational cost and high quality results.

First, a novel real-time high-quality shot detection strategy is proposed. While abrupt transitions are detected through a very fast pixel-based analysis, gradual transitions are obtained from an efficient edge-based analysis. Both analyses are reinforced with a motion analysis that allows to detect and discard false detections. This analysis is carried out exclusively over a reduced amount of candidate transitions, thus maintaining the computational requirements.

On the other hand, a moving object detection strategy, which is based on the popular Mixture of Gaussians method, is proposed. This strategy, taking into account the recent history of each image pixel, adapts dynamically the amount of Gaussians that are required to model its variations. As a result, we improve significantly the computational efficiency with respect to other similar methods and, additionally, we reduce the influence of the used parameters in the results.

Alternatively, in order to improve the quality of the results in complex scenarios containing dynamic backgrounds, we propose different non-parametric based moving object detection strategies that model both background and foreground. To obtain high quality results regardless of the characteristics of the analyzed sequence we dynamically estimate the most adequate bandwidth matrices for the kernels that are used in the background and foreground modeling. Moreover, the application of a particle filter allows to update the spatial information and provides a priori knowledge about the areas to analyze in the following images, enabling an important reduction in the computational requirements and improving the segmentation results. Additionally, we propose the use of an innovative combination of chromaticity and gradients that allows to reduce the influence of shadows and reflects in the detections.

Índice

Índice de figuras

Índice de Tablas

Capítulo 1

Introducción

Comenzar bien no es poco,
pero tampoco es mucho.

Sócrates (470 AC-399 AC),
filósofo griego.

1.1. Motivación

Debido a los importantes avances tecnológicos conseguidos recientemente y al crecimiento de Internet, en los últimos años ha surgido una importante demanda de aplicaciones orientadas al análisis de secuencias de vídeo (Elhabian et al., 2008) (Chaisorn et al., 2009).

Por un lado, el número de usuarios de aplicaciones de edición de vídeo ha sufrido un importante aumento (Peng et al., 2011). Estas aplicaciones incluyen herramientas de indexación y búsqueda de contenidos que, en una primera etapa, deben ser capaces de segmentar las secuencias analizadas en tomas, detectando eficientemente las transiciones que las separan, tanto las abruptas como las graduales (Smeaton et al., 2010). Sin embargo, para proporcionar resultados de calidad, identificando la mayor parte de transiciones existentes y evitando las falsas detecciones, estas herramientas aplican estrategias complejas que no son suficientemente rápidas y que, además, dependen de numerosos umbrales que han de ser adecuadamente establecidos por los usuarios en función de las características de cada secuencia analizada (Brezeale y Cook, 2008).

Por otro lado, han aparecido numerosos dispositivos electrónicos de última generación que incluyen cámaras de vídeo como, por ejemplo, teléfonos inteligentes (*smart-phones*), *tablets*, o vídeo-consolas portátiles y de sobremesa (Sangani, 2010). Las aplicaciones demandadas por todos estos dispositivos han de ser capaces de trabajar a la mayor velocidad posible y, además, han de proporcionar resultados de gran calidad en un amplio número de escenarios con características muy variadas (Sheikh et al., 2009). Estas aplicaciones, para realizar tareas de alto nivel como, por ejemplo, el seguimiento de objetos móviles, su clasificación, o el análisis de eventos, hacen uso de técnicas de detección de objetos móviles (Landabaso y Pardas, 2008). Sin embargo, aunque a día de hoy se han propuesto numerosas estrategias de detección de objetos móviles, ninguna de ellas cumple los requisitos requeridos

por estas aplicaciones: trabajar rápido, proporcionar buenos resultados independientemente de las características del entorno en el que se utilicen y depender del menor número posible de parámetros.

Por lo tanto, como respuesta a esta importante demanda de aplicaciones para la segmentación temporal de secuencias de vídeo y para la detección de los objetos móviles presentes en las mismas, en esta tesis se proponen distintas estrategias que cumplen los requisitos de calidad, velocidad y facilidad de uso requeridos por los usuarios de dichas aplicaciones.

1.2. Contenido de la tesis

Además del capítulo de introducción, la tesis consta de otros seis capítulos y de tres apéndices. A continuación se describe brevemente el contenido de cada uno de ellos.

En el capítulo 2 se presenta el estado del arte correspondiente al conjunto de técnicas utilizadas para llevar a cabo esta tesis. Por un lado se analizan las estrategias más relevantes dentro del campo de la segmentación temporal de secuencias en tomas y, por otro lado, se describen las estrategias propuestas en la literatura para llevar a cabo la segmentación de imágenes y la detección de objetos móviles en secuencias de vídeo.

En el capítulo 3 se describe la estrategia propuesta para localizar las transiciones, tanto abruptas como graduales, que separan las tomas de una secuencia de vídeo. En una primera etapa, dicha estrategia aplica métodos rápidos y eficaces, basados en comparaciones entre píxeles de imágenes consecutivas y en el análisis de los bordes a lo largo de las secuencias, que permiten detectar la inmensa mayoría de las transiciones existentes. En una segunda etapa, mediante un análisis de movimiento realizado sobre las imágenes que delimitan las transiciones resultantes de la primera etapa, se detectan y descartan la mayor parte de las falsas detecciones. El resultado es un sistema de detección de tomas que es capaz de obtener muy buenos resultados y que, además, debido a su bajo coste computacional asociado, puede ser utilizado en aplicaciones que requieran trabajar a gran velocidad.

En el capítulo 4 se presenta una estrategia de detección de objetos móviles, basada en el método de mezcla de gaussianas, capaz de proporcionar resultados de gran calidad en tiempo real. La principal aportación de dicha estrategia es su capacidad para adaptar dinámicamente el número de gaussianas asociadas a cada píxel en cada instante temporal. De ese modo se consigue reducir muy notablemente tanto la carga computacional del método original como sus requisitos de memoria. Además, el algoritmo de detección propuesto reduce la dependencia de los resultados con los parámetros utilizados por el método original, facilitando así su utilización independientemente de las características de la secuencia sobre las que se aplique.

En el capítulo 5 se describe otro sistema de detección de objetos móviles, en este caso basado en técnicas de modelado no paramétrico, que es capaz de proporcionar resultados de mejor calidad que la estrategia descrita en el capítulo 4, especialmente en las secuencias en las que las variaciones de los píxeles no pueden modelarse con métodos paramétricos. Esta estrategia, a diferencia de otros métodos de detección, modela tanto el fondo como el primer plano de la secuencia y, además de utilizar información de apariencia de los

píxeles (color, gradiente, etc.), utiliza información relativa a su posición dentro de la imagen. Adicionalmente, aplica dos estrategias que permiten estimar dinámicamente el ancho de banda de las funciones locales utilizadas en ambos modelados. El resultado es una estrategia de detección que consigue resultados de gran calidad incluso en situaciones en las que el fondo y el primer plano son parecidos y que, además, siempre utiliza funciones locales con un ancho adecuado a las características de la secuencia analizada. Por otro lado, gracias a la aplicación de una innovadora estrategia de seguimiento basada en un filtro de partículas, la posición de los objetos móviles previamente detectados se actualiza de imagen a imagen, lo cual resulta en una apreciable mejora de la calidad de los resultados y en una significativa reducción del coste computacional asociado al modelado del primer plano.

En el capítulo 6 se describen algunas estrategias de mejora para aplicar al sistema de detección descrito en el capítulo 5. Por un lado se propone utilizar un innovador conjunto de características de apariencia de los píxeles, compuesto por sus componentes de color normalizado y por el módulo del gradiente de su valor de saturación. De esta forma, a diferencia de los sistemas de detección que utilizan las componentes de color RGB, se consigue descartar la mayor parte de las falsas detecciones debidas a las sombras y a los reflejos que provocan los objetos móviles y que, en sistemas como los descritos en los capítulos anteriores, son una importante limitación Por otro lado se plantean dos estrategias que permiten mejorar muy significativamente la eficiencia computacional de los métodos de detección basados en el modelado no paramétrico de fondo y primer plano ya que, posiblemente, el elevado coste computacional de dichos métodos sea su mayor limitación frente a otros tipos de estrategias de detección. En primer lugar se propone una estrategia que evita la necesidad de utilizar la información espacial de los píxeles en el modelado del fondo y, así, reduce muy significativamente el coste computacional asociado a dicho modelado. En segundo lugar se utiliza una novedosa estrategia que, a partir de la información proporcionada por el filtro de partículas previamente mencionado, es capaz de determinar qué regiones de la imagen deben ser analizadas en cada instante, evitando el análisis de un elevado porcentaje de los píxeles.

En el capítulo 7 se exponen las conclusiones extraídas de la tesis y se plantean algunas opciones de trabajo futuro.

En el apéndice A se describen las bases de datos utilizadas para probar la calidad de los resultados proporcionados por las estrategias propuestas a lo largo de la tesis. En primer lugar se describe la base de datos utilizada para evaluar la calidad de la estrategia de detección de tomas descrita en el capítulo 3. En segundo lugar se describe la base de datos utilizada para valorar la calidad de las estrategias de detección de objetos móviles presentadas en los capítulos 4, 5 y 6.

En el apéndice B se analiza el modo en el que es posible relacionar la probabilidad acumulada de una distribución gaussiana multidimensional con la distancia a su centro. Primero se describe una transformación afín que permite convertir una gaussiana cualquiera en otra gaussiana estándar, la cual puede a su vez relacionarse con una distribución de tipo *chi-cuadrado*, en la que el análisis de la relación entre distancia y probabilidad acumulada es fácil de realizar. Por último se analiza el caso concreto de una distribución gaussiana bidimensional, ya que dicho caso es aplicado varias veces a lo largo de la tesis.

En el apéndice C se describe el filtro de partículas utilizado en las estrategias descritas en los capítulos 5 y 6 para actualizar las posiciones de los objetos móviles previamente detectados. A diferencia de otros filtros, el que nosotros proponemos es capaz de trabajar eficientemente con un número variable de regiones móviles. En cada instante, las partículas se reparten entre las regiones móviles existentes y se realizan estimaciones independientes para cada una de ellas. Además, incluye una estrategia de análisis de regiones que permite identificar si alguna de las regiones no está siendo adecuadamente representada por las partículas y, así, detecta la presencia de nuevas regiones.

1.3. Contribuciones científicas

A continuación se describen brevemente las publicaciones científicas que han derivado de esta tesis y se especifica con qué capítulos o secciones de la misma están relacionadas.

- En (Cuevas y García, 2010b) se describe una primera versión del sistema de detección de transiciones descrito en el capítulo 3.
- En (Cuevas et al., 2008) aparece la estrategia de detección de objetos móviles descrita en el capítulo 4.
- En (Nieto et al., 2010) se describe una primera versión del filtro de partículas descrito en el apéndice C, aplicado sobre las detecciones obtenidas mediante la estrategia descrita en el capítulo 4.
- En (Cuevas et al., 2010a) se propone una estrategia basada en la realimentación entre detección y seguimiento en la que, al igual que los sistemas de detección descritos en los capítulos 5 y 6, los algoritmos de detección empleados mejoran sus resultados gracias a la información proporcionada por una estrategia de seguimiento que, a su vez, haciendo uso de las detecciones es capaz de ofrecer trayectorias más precisas.
- En (Cuevas y García, 2010c) se describe una estrategia de detección basada en el modelado no paramétrico del primer plano y del fondo en la que, como alternativa las comúnmente utilizadas componentes RGB, para reducir las falsas detecciones debidas a sombras y reflejos se propone la utilización de componentes de color normalizadas y de gradientes (sección 6.1 del capítulo 6).
- En (Cuevas et al., 2010b) se propone una estrategia de detección que, suprimiendo el uso de la información espacial en el modelado del fondo (sección 6.3 del capítulo 6), es capaz de mejorar muy notablemente la eficiencia computacional de otras estrategias que, de forma similar, modelan tanto el fondo como el primer plano.
- En (Cuevas y García, 2010a) se describe un esquema de detección en el que, además de utilizarse el conjunto de características propuestas para evitar las falsas detecciones debidas a sombras y reflejos (sección 6.1 del capítulo 6), se hace uso del filtro de partículas descrito en el apéndice C. Además, se propone la utilización de las predicciones proporcionadas por dicho filtro para: obtener información a priori que permita mejorar la calidad de los resultados (sección 5.5.2 del capítulo 5); y reducir el número de píxeles analizados en cada instante (sección 6.4 del capítulo 6).
- En (Cuevas y García, 2011) se describen dos estrategias de estimación dinámica del

ancho de las funciones locales utilizadas en el modelado del fondo (sección 5.23 del capítulo 5) y del primer plano (sección 5.28 del capítulo 5), aplicadas a un esquema de detección en el que sólo se utiliza información espacial para modelar el primer plano.

Capítulo 2

Estado del arte

No hay que confundir nunca el conocimiento con la sabiduría. El primero nos sirve para ganarnos la vida; la sabiduría nos ayuda a vivir.

Sorcha Carey (1943-¿?),
profesora de arte clásico inglés.

Resumen: En este capítulo se presenta el estado del arte correspondiente al conjunto de técnicas utilizadas para el desarrollo de la presente tesis, las cuales pueden englobarse dentro de dos grandes campos. Por un lado, la sección 2.1, se centra en el campo correspondiente a la segmentación temporal de secuencias de vídeo. En esta sección se describen tanto los objetivos de la segmentación temporal, como las etapas de las que consta y las soluciones propuestas en la literatura, a lo largo de los últimos años, para llevarla a cabo exitosamente. Por otro lado, en la sección 2.2, se presenta el estado del arte correspondiente al campo de la segmentación de objetos en imágenes. Dentro de esta sección se ofrece una descripción detallada de los objetivos de dicho campo, de las dificultades que presenta y de las propuestas llevadas a cabo para obtener resultados de calidad en el mayor número posible de situaciones.

2.1. Segmentación temporal de secuencias de vídeo

A lo largo de los últimos años, junto con el crecimiento de Internet y de las mejoras informáticas (ordenadores cada vez más rápidos y de mayor capacidad), han aparecido numerosos avances relacionados con la interpretación de información multimedia (Chaisorn et al., 2009).

Estos avances han hecho necesaria la aparición de aplicaciones capaces de trabajar con información multimedia que incluye vídeos de gran tamaño. Algunas de estas aplicaciones son, por ejemplo: las librerías digitales, la enseñanza a distancia, el vídeo bajo demanda, la transmisión de vídeo digital, la televisión interactiva, o los sistemas de información multimedia. Para posibilitar la gestión de toda esta información se han desarrollado herramientas

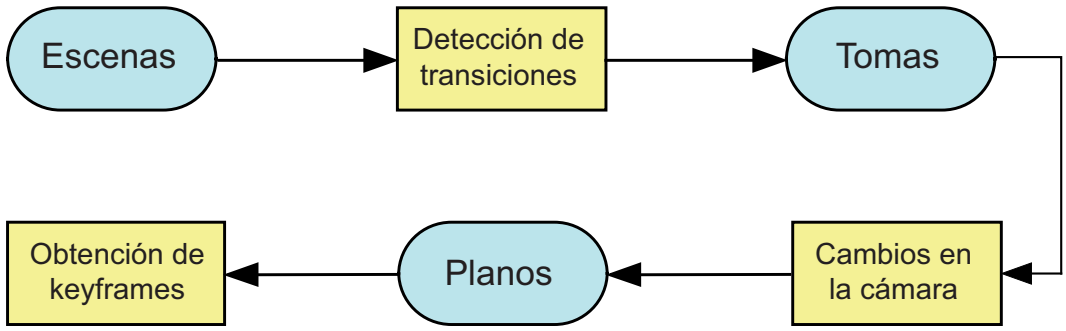


Figura 2.1: Etapas típicas en la segmentación temporal de secuencias de vídeo.

capaces de realizar operaciones de indexación, de búsqueda de contenidos y de recuperación de información relevante dentro de este material. La aparición de estas herramientas ha sido posible gracias a las numerosas estrategias de segmentación temporal de secuencias de vídeo desarrolladas durante los últimos años (Lew et al., 2006).

El objetivo principal de la segmentación de vídeo es la extracción automática de la información referente a su contenido. Para conseguir esta información se lleva a cabo una segmentación temporal de las secuencias o, lo que es lo mismo, su división o estructuración en unidades homogéneas desde algún punto de vista (luminosidad media, distribución de color, movimiento de la cámara, etc.) (Bescós, 2001). De este modo es posible la elaboración de un índice del contenido de las secuencias, mediante el cual se puede acceder a los distintos tramos en los que éstas han sido divididas (motivo por el cual las técnicas de segmentación temporal también son conocidas como técnicas de indexación de vídeo).

2.1.1. Etapas presentes en la segmentación de secuencias

La figura 2.1 muestra un esquema de segmentación de secuencias en el que las secuencias son divididas a partir de la localización de determinados eventos temporales (transiciones y variaciones de la cámara). La estructuración temporal que se hace en esta figura está hecha desde el punto de vista de la producción y la post-producción de vídeo, y divide las secuencias en escenas, tomas y planos (Bescós, 2001). A continuación se incluye una breve descripción de cada una de estas unidades temporales:

- **Escena**: Intervalo de la secuencia a lo largo del cual todas las imágenes tienen relación con un mismo objeto o grupo de objetos. Una escena puede ser a su vez dividida en varias tomas.
- **Toma**: Segmento ininterrumpido de una escena de vídeo o, lo que es lo mismo, un conjunto de imágenes consecutivas que resultan de una única y continuada operación de grabación de la cámara. Las transiciones entre tomas son creadas en el proceso de post-producción, utilizando técnicas de edición de vídeo. Estas transiciones pueden ser de dos tipos: abruptas o graduales. De la misma forma que las escenas, las tomas pueden ser divididas en segmentos de menor tamaño. Estos segmentos son denominados planos.
- **Plano**: Conjunto de imágenes contenido en una toma y caracterizado por la situación

Figura 2.2: Ejemplos de transiciones entre tomas. (a) Transición abrupta. (b) *Dissolve*. (c) *Fade-in*. (d) *Fade-out*. (e) *Wipe*.

relativa entre el escenario y la cámara. Por lo tanto, las transiciones entre planos vendrán dadas por cambios en la posición u orientación de la cámara y por los cambios en la distancia focal de la misma.

Una vez ha sido realizada la división de una secuencia en escenas, tomas y planos, se procede al análisis de cada uno de los segmentos obtenidos por separado, para caracterizar el contenido de los mismos y así extraer una o varias imágenes representativas de cada intervalo. Estas imágenes representativas de cada segmento son las denominadas "*keyframes*". A partir de estas imágenes características se pueden realizar estudios sencillos (análisis de color, de forma, de movimiento, etc.), o más complejos (detección y seguimiento de objetos, detección de personas, etc.).

2.1.2. Transiciones abruptas y graduales

Tal y como se ha mencionado anteriormente, las transiciones entre tomas pueden ser de dos tipos: abruptas o graduales.

Figura 2.3: Ejemplos de variaciones en la cámara. (a) *Pan*. (b) *Tilt*. (c) *Zoom*. (d) *Traveling*.

Las transiciones abruptas se localizan en instantes temporales en los que dos imágenes consecutivas pertenecen a dos tomas distintas, provocando de este modo una clara discontinuidad en la secuencia. En la figura 2.2.a se muestra un ejemplo de este tipo de transición. En este ejemplo, formado por cuatro imágenes consecutivas, se ve claramente como entre la segunda y la tercera imagen se ha producido un cambio de cámara que da lugar a una transición abrupta.

Por otro lado, las transiciones graduales resultan de la aplicación de efectos de edición de vídeo más complejos, en los que se ven involucradas más de dos imágenes. Estas transiciones surgen de la aplicación de una transformación gradual o progresiva que se origina en la última imagen de la toma anterior a la transición y finaliza en la primera imagen de la toma posterior a la transición. Las transiciones graduales pueden ser de varios tipos:

- ***Dissolve* (fundido)**: Fusión entre dos imágenes pertenecientes a dos tomas distintas. Una imagen aparece gradualmente, mientras la otra va desapareciendo. La figura 2.2.b muestra un ejemplo de transición gradual de tipo *dissolve*.
- ***Fade-in***: Crecimiento gradual de la intensidad de las imágenes, partiendo de una imagen negra. En la figura 2.2.c aparece representada una transición de este tipo.
- ***Fade-out***: Decrecimiento gradual del brillo de las imágenes hasta convertirse en una imagen negra. En la figura 2.2.d se muestra un ejemplo de este tipo de transición.

- ***Wipe* (cortinilla)**: Una toma aparece gradualmente, superponiéndose a la anterior (por ejemplo, del modo mostrado en la figura 2.2.e).

2.1.3. Variaciones de la cámara

Las variaciones que se producen en la cámara y que hacen que cambie su posición con respecto a la del escenario, son las que determinan los cambios de plano dentro de una misma toma. Estas variaciones pueden ser de varios tipos:

- ***Pan***: Movimiento de la cámara en su eje vertical, lo que se traduce en un desplazamiento horizontal de las imágenes capturadas (se puede ver un ejemplo en la figura 2.3.a).
- ***Tilt***: Movimiento de la cámara en su eje horizontal, lo que se traduce en un desplazamiento vertical de las imágenes capturadas (se puede ver un ejemplo en la figura 2.3.b).
- ***Zoom***: Variación de la distancia focal de la cámara (se puede ver un ejemplo en la figura 2.3.c).
- ***Traveling***: Desplazamiento de la cámara (se puede ver un ejemplo en la figura 2.3.d).

2.1.4. Técnicas para la detección de cambios de toma

Las propuestas centradas en la localización de transiciones abruptas y graduales se vienen realizando desde aproximadamente los últimos 25 años. En estos trabajos se ha desarrollado un importante número de estrategias relevantes, pudiéndose encontrar varios informes recopilatorios que describen globalmente la situación y los avances en este campo de investigación (Lefèvre et al., 2003) (Snoek y Worring, 2005) (Lew et al., 2006) (Truong y Venkatesh, 2007) (Brezeale y Cook, 2008) (Seidl et al., 2010).

En los primeros estudios realizados se prestaba más atención a la detección de transiciones abruptas, pero a medida que éstas eran localizadas de forma más eficiente se fue dando más importancia a la detección de transiciones graduales, cuya identificación resulta de mayor complejidad y dificultad debido a los múltiples tipos de transiciones existentes.

En términos generales, la detección de cambios de toma consiste en la localización de diferencias apreciables entre imágenes consecutivas dentro de una misma escena. Por ejemplo, si entre dos imágenes consecutivas se aprecia una diferencia grande, se puede asumir que entre dichas imágenes existe una transición abrupta. Si, por el contrario, en vez de diferencias grandes, se aprecian diferencias crecientes a lo largo de sucesivas imágenes, se puede asumir que dichas imágenes formarán parte de una transición gradual.

En esta sección se presenta una breve descripción de las técnicas de detección de transiciones, tanto abruptas como graduales, que han sido fruto de la investigación realizada los últimos años en el campo de la segmentación temporal de secuencias.

2.1.4.1. Técnicas para la detección de transiciones abruptas

Las transiciones abruptas han sido las más ampliamente estudiadas, y son muchas las técnicas capaces de detectarlas correctamente. Estas técnicas normalmente se centran en el

análisis de secuencias con características comunes; es decir, un tipo de secuencias concreto (por ejemplo, deportes o noticias). Sin embargo, las técnicas de análisis que pretenden segmentar un conjunto más variado de tipos de secuencias (películas, series, anuncios publicitarios, dibujos animados, noticias, documentales, deportes, etc.) no logran detecciones tan satisfactorias, incrementando tanto el número de falsas detecciones como el de transiciones no detectadas.

De todas estas técnicas, las más sencillas e intuitivas basan su funcionamiento en la realización de comparaciones entre imágenes consecutivas a nivel de píxel. Uno de los primeros métodos que aparecen en la literatura es el presentado en (Nagasaka y Tanaka, 1992), el cual detecta las transiciones abruptas a partir del análisis de la suma de los valores absolutos de las diferencias entre cada par de píxeles, situados en la misma posición espacial, en imágenes consecutivas. En el caso de que esta diferencia supere el valor de un umbral predefinido se determina que existe una transición abrupta.

Posteriormente, se propusieron algunas técnicas que toman como base a la anterior. Este es el caso de la estrategia propuesta en (Zhang et al., 1993). Dicho trabajo, propone la detección de transiciones abruptas mediante la aplicación de dos umbrales: uno para decidir si un píxel ha cambiado o no, y otro para determinar si han cambiado suficientes píxeles como para determinar la existencia de una transición. Además, para reducir el número de falsas detecciones debidas a los movimientos globales de la cámara y a los de los objetos móviles de gran tamaño, no compara directamente los valores de los píxeles, sino el resultado de un filtrado paso bajo (considerando ventanas de 3×3 píxeles) aplicado sobre los mismos. A parte de éstas estrategias, también se han propuesto otros trabajos que plantean la utilización de diferentes medidas estadísticas a nivel de píxel para la localización de transiciones abruptas (Ren et al., 2001). Todas estas técnicas, a pesar de ser sencillas, tienen la desventaja de no ofrecer resultados suficientemente buenos, dando lugar a un elevado número de falsos positivos y pasando por alto muchas transiciones abruptas.

Más recientemente, para tratar de reducir la cantidad de falsas detecciones debidas a la existencia de movimientos globales, aparecieron técnicas basadas en el análisis estadístico de los píxeles (Yuan y Feng, 2010) (Kucuktunc et al., 2010). En estas estrategias, si el valor resultante de evaluar la diferencia entre histogramas u otros estadísticos de alguna de las características de las imágenes es superior a un determinado umbral, se determina que existe una transición abrupta (Lefèvre et al., 2003). Estos análisis son una evolución de las técnicas presentadas en trabajos como el de (Tonomura y Abe, 1990), en el que únicamente se utiliza la información correspondiente a la componente de luminancia de los píxeles.

Por otro lado, como alternativa a las técnicas basadas en las comparaciones a nivel de píxel y a las basadas en histogramas, aparecieron estrategias más robustas que hacen uso de la información obtenida a partir de diferentes espacios de color (Pye et al., 1998) (Ahmed et al., 1999). Estas estrategias, a lo largo de los últimos años, han ido incorporando medidas estadísticas más complejas (Ionescu et al., 2006) (Camara-Chavez et al., 2006) (Chasanis et al., 2009), siendo capaces de proporcionar resultados de más calidad que las anteriores y con un bajo coste computacional asociado (eso sí, siempre trabajando sobre familias de secuencias de características similares).

Debido a que gran parte de las falsas transiciones se deben a los movimientos de la

cámara o a los de los objetos móviles, también se han propuesto numerosas técnicas basadas en el análisis del movimiento a lo largo de las secuencias (Bouthemy et al., 1999). Algunas de estas estrategias aplican técnicas de compensación del movimiento antes de llevar a cabo el análisis de los píxeles (Liu et al., 2002), mientras que otras analizan el desplazamiento de puntos característicos identificados en las imágenes (Li et al., 2010). Estas técnicas, aunque en algunas situaciones proporcionan resultados de gran calidad, tienen asociado un elevado coste computacional.

Otra posibilidad muy común, en el caso de trabajar con secuencias comprimidas, es el análisis de éstas en el dominio comprimido. De esta forma se evita el proceso de descompresión, el cual siempre es muy costoso y, además, se aprovecha la información que contiene la secuencia en su versión codificada. Algunos de los trabajos que se centran en el análisis de secuencias comprimidas son los presentados en (Zhang et al., 1995) (Lee et al., 2002) (De Bruyne et al., 2006) (Cao y Cai, 2007) (Lin et al., 2011). Estos trabajos generalmente explotan características que son fácilmente extraíbles en el dominio comprimido, como pueden ser: la información de los cuadros I, los vectores de movimiento, la relación entre el número de macrobloques con predicción hacia delante o hacia atrás y los coeficientes DC. Los resultados obtenidos con estos métodos, teniendo en cuenta algunos análisis como el realizado en (Koprinska y Carrato, 2001), son peores que los obtenidos con los métodos expuestos previamente, a pesar de ofrecer la ventaja de no necesitar descomprimir la señal.

Por último, otra de las alternativas utilizadas en otros trabajos es la que consiste en hacer uso de diferentes técnicas de forma simultánea (Lee et al., 2000) (Sánchez y Binefa, 2003) (Ciocca, 2010). De ese modo, al aprovechar las ventajas de cada una de las técnicas que combinan, se consiguen resultados de gran calidad. Sin embargo, la combinación de varias técnicas que por sí solas ya son computacionalmente costosas hace que estas estrategias no sean apropiadas para su utilización en sistemas que requieran trabajar en tiempo real.

2.1.4.2. Técnicas para la detección de transiciones graduales

A pesar de existir un gran número de trabajos relacionados con la detección de transiciones graduales, ninguno de ellos es capaz de obtener resultados suficientemente buenos y ninguno parece ofrecer resultados significativamente mejores que el resto (Chasanis et al., 2009). El principal problema en la detección de este tipo de transiciones, además de la gran variedad de transiciones existentes, es la aparición de zonas de la escena con mucho movimiento (ya sea debido a movimientos de la cámara o a la aparición en la escena de objetos móviles de gran tamaño).

Debido a su falta de rigurosidad (no proporcionan resultados con datos objetivos) y al tamaño y representación de las muestras utilizadas, resulta difícil evaluar la mayor parte de estos trabajos (Brezeale y Cook, 2008). En cualquier caso, la calidad de los resultados que obtienen es muy inferior a la alcanzada por los métodos previamente descritos para la detección de transiciones abruptas.

Algunos de los métodos de detección de transiciones abruptas, a los que ya se ha hecho referencia en sección 2.1.4.1, también se han aplicado a la detección de transiciones graduales. Esto es debido a que, en definitiva, la existencia de transiciones graduales se determina mediante la aplicación de un umbral al resultado de una medida de diferencia, y la reduc-

ción del valor de este umbral permitirá la detección de cambios más suaves (transiciones graduales), aunque a costa de la aparición de un alto número de falsos positivos. El principal problema de estas técnicas es la aparición de movimiento global, el cual puede causar el mismo efecto que una transición gradual, haciendo que el número de falsas detecciones sea muy elevado (Lin et al., 2008) (Plotkowiak y Lay, 2008).

Los resultados obtenidos mediante este conjunto de técnicas generales pusieron de manifiesto las grandes dificultades a la hora de establecer un método de detección válido para cualquier tipo de transición gradual. Por este motivo, muchos de los trabajos posteriores establecen modelos específicos para cada uno de los tipos de transiciones que se desea detectar, principalmente *dissolves* y *wipes* y, por este motivo, se han desarrollado detectores adaptados a dichos modelos. Así, el trabajo presentado en (Kwon et al., 2008) basa la detección de un *dissolve* en el análisis de la desaparición gradual de los bordes correspondientes a la toma anterior y la aparición de nuevos bordes pertenecientes a la siguiente toma. Al igual que este, existen otros muchos trabajos basados en el análisis de los bordes a lo largo de las secuencias, (Hauptmann et al., 2003) (Lee et al., 2003) (Huan et al., 2008), o en el de puntos singulares identificados en las imágenes (Wang et al., 2009) (Fu y Zeng, 2009).

En cuanto a la detección de cortinillas o *wipes*, uno de los trabajos más representativos es el presentado en (Fernando et al., 2002). Este trabajo está basado en la detección de la línea recta que barre la secuencia de imágenes (la cortinilla en sí) y se basa en el cálculo de estadísticos de luminancia, que son calculados dividendo las imágenes en bloques. La decisión final sobre si hay o no hay una recta se realiza utilizando la Transformada de *Hough* (Ballard, 1981). Además de este trabajo, también es posible encontrar algunos otros más actuales (Yufeng et al., 2010) (Warhade et al., 2010). Sin embargo, dado que las cortinillas son muy poco frecuentes en la mayor parte de las secuencias de vídeo, la cantidad de métodos propuestos para su detección es notablemente inferior a la de las estrategias propuestas para detectar las transiciones graduales de tipo *dissolve*.

Además de estas estrategias de detección, del mismo modo que ocurría en el caso de las técnicas de detección de transiciones abruptas, también han aparecido trabajos que tratan de detectar la presencia de transiciones graduales analizando las secuencias en el dominio del espacio y de la frecuencia (Drew et al., 2002) (Yao et al., 2008), y trabajos que operan en el dominio comprimido (Damghanian et al., 2006) (Ren et al., 2010). Además, también es posible encontrar estrategias combinadas que analizan soluciones espacio-temporales obtenidas a partir de imágenes DC (Joyce y Liu, 2006), o propuestas que realizan análisis estadísticos de distintos tipos de información (Padalkar y Zaveri, 2010).

Independientemente del mayor o menor grado de acierto obtenido por estos métodos, la gran desventaja que tienen radica en que están orientados a la detección de *dissolves* y *wipes* de progresión lineal. Aunque este tipo de transiciones fueron las primeras en aparecer, hace tiempo que las técnicas de edición no lineal de vídeo abrieron las puertas a un gran número de efectos de transición. Ante esta situación el uso de detectores centrados en características muy específicas de cada tipo de transición resulta poco viable (Seidl et al., 2010).

2.1.5. Técnicas para la detección de cambios de plano

Los cambios de plano están determinados por la aparición de variaciones de la cámara respecto a la escena (*pan*, *tilt*, *zoom* o *traveling*). A menudo, estos cambios producidos en la cámara dan lugar a falsas clasificaciones, ya que en la mayor parte de los métodos de localización de transiciones graduales la presencia de movimiento en la cámara es confundida con efectos de *dissolve* (Koprinska y Carrato, 2001). Sin embargo, en el contexto de los sistemas de recuperación de vídeo, que necesitan la selección de *keyframes*, la extracción de índices, o la búsqueda de contenidos, la identificación de los movimientos de la cámara es de gran importancia (Duan et al., 2006). Es por este motivo por el que han aparecido numerosos trabajos que desarrollan técnicas de detección de cambios en la cámara de adquisición.

Los movimientos de la cámara dan lugar a la aparición de campos de vectores de movimiento característicos. A causa de esto, muchos de los trabajos existentes están basados en el análisis de los vectores de movimiento que aparecen a lo largo de las secuencias (Micheloni et al., 2010). Partiendo de este tipo de análisis, a lo largo de los últimos años se han propuesto distintas estrategias que, aplicando determinadas reglas para analizar la dirección de los vectores de movimiento, permiten detectar el *pan*, el *tilt*, o el *zoom* de la cámara (Jeon et al., 2005) (Jiayin et al., 2010). Para tal propósito, estos trabajos parten de las siguientes hipótesis: cuando se produce un *pan* o un *tilt*, la mayor parte de los vectores de movimiento son paralelos a un vector medio que identifica el movimiento de la cámara; y en el caso de un *zoom*, los vectores de movimiento mostrarán un foco de expansión (si la distancia focal ha aumentado), o un foco de contracción (si la distancia focal se ha reducido).

Otros trabajos, como los presentados en (Dumitras y Haskell, 2004) (Suhr et al., 2011), en vez de analizar el movimiento global en las imágenes, tratan de caracterizar los movimientos o cambios de la cámara analizando los desplazamientos en distintas regiones de las imágenes. De ese modo tratan de identificar si los cambios detectados afectan a toda la imagen y, así, reducen la aparición de falsas detecciones debidas a objetos móviles de gran tamaño que, aunque afectan a gran parte de las imágenes, no afectan a muchas de las regiones analizadas por separado. Estas estrategias ofrecen buenos resultados, pero suelen ser muy sensibles al ruido y tienen asociado un elevado coste computacional.

Otra de las alternativas propuestas para la detección de cambios de plano es el análisis de las secuencias en el dominio comprimido (Duan et al., 2006) (Tao et al., 2009) (Ren et al., 2010). La ventaja de estas estrategias es que disponen de los campos de vectores de movimiento que proporcionan los algoritmos de compresión y, por lo tanto, se ahorran el elevado coste computacional que supone su obtención. Sin embargo, dado que estos vectores son aquellos que minimizan las diferencias entre bloques de las imágenes comparadas, no siempre son una buena representación del movimiento real en las imágenes y, debido a eso, es fácil que den lugar a falsas detecciones.

2.2. Segmentación de objetos

La segmentación de objetos es una de las principales herramientas en el tratamiento digital de imágenes, resultando de gran ayuda tanto en aplicaciones que analizan imágenes sueltas

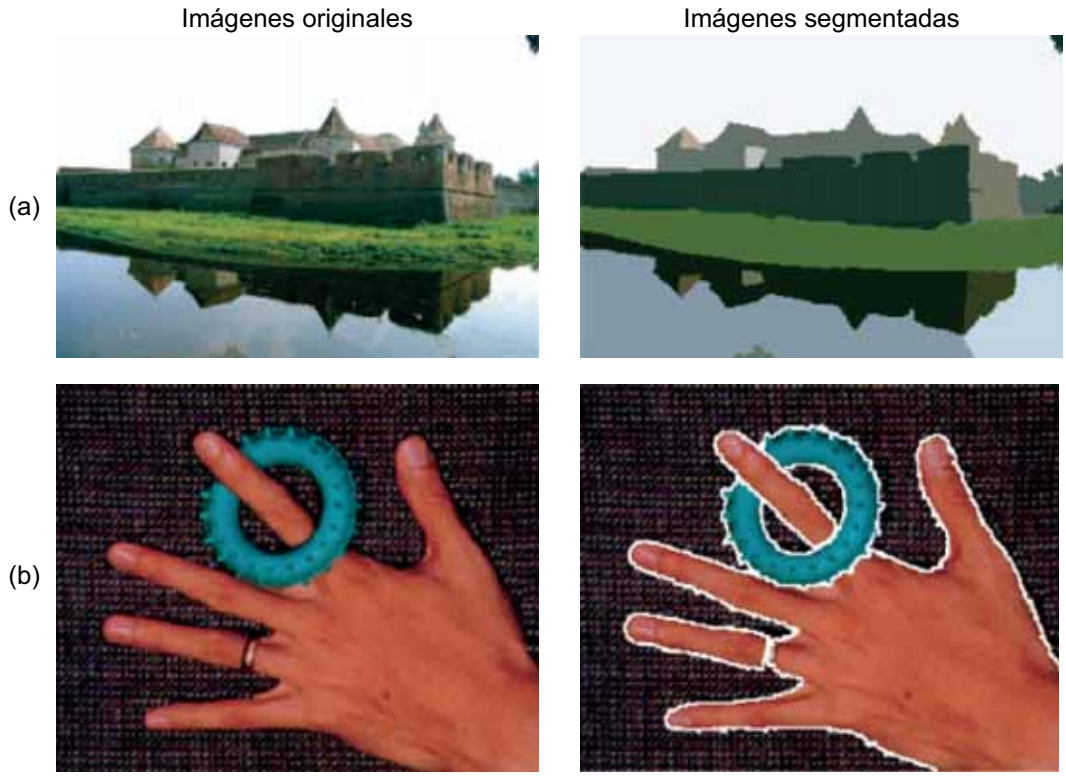


Figura 2.4: (a) Ejemplo de segmentación basada en colores. (b) Ejemplo de segmentación basada en contornos.

como en aquellas que trabajan sobre secuencias de vídeo (Cufi et al., 2003):

- Permite simplificar la adquisición de medidas.
- Permite obtener un análisis cuantitativo del contenido de las imágenes.
- Facilita la reconstrucción de imágenes.
- Hace posible la extracción de información de zonas de las imágenes ocultas al ojo humano.
- Permite la realización de tareas repetitivas.

Desde el punto de vista del procesamiento de imágenes, la segmentación puede definirse como un proceso de partición en el que una imagen es dividida en múltiples regiones constituyentes no solapadas (Coto, 2005), las cuales son homogéneas con respecto a alguna característica (color, intensidad, textura, etc.). Por otro lado, las regiones adyacentes deberán ser lo suficientemente distintas en cuanto a esta característica se refiere. El objetivo de la segmentación es simplificar y/o modificar las imágenes, de forma que se facilite el análisis y la extracción de los objetos representativos contenidos en las mismas (Shapiro y Stockman, 2001).

De este modo, el proceso de segmentación de una imagen cualquiera I consiste en determinar el conjunto de regiones, $S_k \subset I$, tal que su unión constituya la imagen I completa.

Por lo tanto, dicho conjunto debe satisfacer que:

$$I = \bigcup_{k=1}^{K} S_k \quad / \quad S_k \bigcap S_j = \phi \quad \forall k \neq j \tag{2.1}$$

En la figura 2.4 se muestran dos ejemplos de imágenes segmentadas. El primer ejemplo, figura 2.4.a, presenta el resultado de un proceso de segmentación basado en el análisis de los colores de la imagen original. La imagen segmentada correspondiente a este ejemplo muestra el conjunto de regiones obtenidas, donde cada región ha sido representada por el valor medio del color de todos los píxeles que la conforman. El segundo ejemplo, figura 2.4.b, presenta una segmentación basada en el análisis de las texturas. En este caso, en el resultado se han representado los contornos que delimitan las distintas texturas conexas identificadas en la imagen original.

Algunas aplicaciones en las que la segmentación de imágenes ocupa un importante papel son, por ejemplo (Varshney et al., 2010):

- El análisis de imágenes médicas.
- La localización de objetos en imágenes adquiridas desde satélites.
- El reconocimiento de rostros.
- Los sistemas de control automático de tráfico.
- Los sistemas de vídeo-vigilancia.

Dependiendo de si las técnicas utilizadas para segmentar utilizan información temporal de la secuencia (información de más de una imagen), o se basan únicamente en la información presente en cada una de las imágenes por separado, las técnicas de segmentación pueden clasificarse en: técnicas a nivel de secuencias, en el primero de los casos; y técnicas a nivel de imágenes, en el segundo.

2.2.1. Segmentación a nivel de imágenes

La segmentación de objetos es un paso clave en las aplicaciones basadas en el análisis de imágenes, ya que facilita la realización de etapas posteriores de más alto nivel como, por ejemplo, la clasificación de imágenes en función de su contenido, la compresión de imágenes, la búsqueda de contenidos en bases de datos, etc.

A día de hoy son muchas las aplicaciones que hacen uso de técnicas de segmentación de imágenes. Sin embargo, la calidad de los resultados que proporcionan está influenciada por algunas características de las imágenes que pueden dar lugar a segmentaciones erróneas. Algunas de estas características poco deseables son, por ejemplo (Kekre et al., 2009):

- **El ruido**: Distorsiona las características de las regiones, haciendo más difícil su análisis.
- **La iluminación**: Un mismo objeto, dependiendo de como esté iluminado y de las sombras que genere, puede dar lugar a la obtención de distintos segmentos.
- **Las formas**: Los objetos con formas complejas complican la determinación de las fronteras que los delimitan.

En la literatura existen numerosas propuestas para segmentar imágenes, las cuales tratan de solventar exitosamente las dificultades que acaban de ser descritas mediante la aplicación de distintas estrategias. Estas estrategias pueden clasificarse en los siguientes grupos (Varshney et al., 2010):

- **Métodos basados en bordes o fronteras** (Hsiao et al., 2006) (Senthilkumaran y Rajesh, 2009) (Wang y Oliensis, 2010): Basan su funcionamiento en la suposición de que existe una correspondencia entre las discontinuidades de intensidad presentes en una imagen, y la frontera de los objetos contenidos en la misma.
- **Métodos basados en regiones**: Utilizan distintos criterios de conectividad entre píxeles para agruparlos en distintas regiones que, muy probablemente, se correspondan con los distintos objetos presentes en la imagen o con un segmento representativo de éstos. Dentro de estas estrategias se puede distinguir entre dos métodos:
  - *Split and merge* (Kelkar y Gupta, 2008) (Chaudhuri y Agrawal, 2010): Técnica que en primer lugar divide las imágenes en pequeños segmentos para, en un segundo paso, agrupar los que satisfagan ciertas condiciones.
  - *Region growing* (Yu y Clausi, 2007) (Yu y Clausi, 2008): Técnica de agrupación de regiones en la que a regiones iniciales, formadas por un único píxel, se les van añadiendo los píxeles contiguos que satisfagan ciertas condiciones establecidas.
- **Métodos basados en técnicas de agrupación (*clustering*)** (Tseng et al., 2006) (Shah et al., 2007) (Ramos y Muge, 2010) (Saha y Maulik, 2011): Tratan de agrupar los píxeles en distintas regiones, a partir de algún criterio de clasificación estadístico y atendiendo a distintas características de los píxeles (color, texturas, etc.).
- **Métodos basados en umbrales** (Maitra y Chatterjee, 2008) (Pérez y González, 2009) (Sathya y Kayalvizhi, 2011): Mediante la aplicación de umbrales, reducen el número de posibles niveles para los píxeles (en cuanto a color, textura, gradientes o cualquier otra característica de los mismos se refiere). De este modo, cada región estará constituida por todos los píxeles que hayan sido clasificados en el mismo nivel.

2.2.2. Segmentación de objetos a nivel de secuencias

En los últimos años, gracias a la gran evolución de los ordenadores, han aparecido muchas aplicaciones que hacen uso de técnicas de procesamiento de imágenes (Wang et al., 2007) como, por ejemplo:

- La vídeo-vigilancia.
- La monitorización.
- El análisis de objetos en movimiento.
- La interacción hombre-máquina.
- La codificación de vídeo basada en objetos (MPEG-4).

En todas estas aplicaciones, la segmentación de las imágenes para la detección de los objetos móviles presentes en las mismas es una etapa fundamental, siendo necesaria para la realización de tareas de más alto nivel como, por ejemplo (Mittal y Paragios, 2004):

- La detección de objetos.

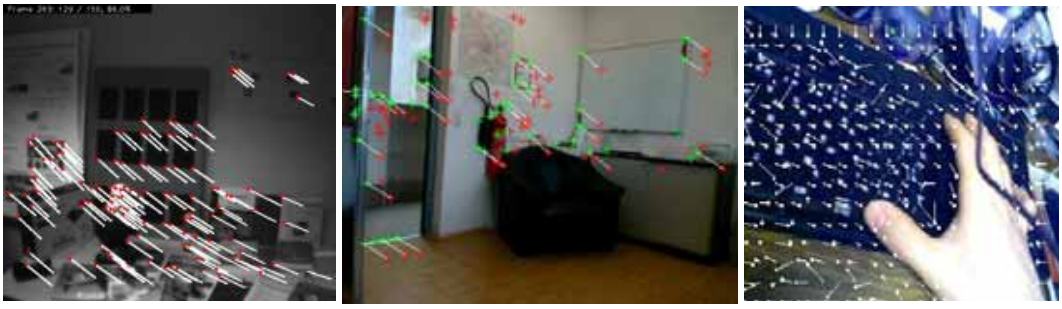


Figura 2.5: Vectores de flujo óptico obtenidos en el proceso de segmentación de secuencias de vídeo. (a) y (b) Vectores resultantes del análisis de puntos singulares. (c) Campo denso de vectores (repartidos uniformemente por toda la imagen).

- El seguimiento de objetos móviles.
- La clasificación de contenidos.
- La detección de eventos.

Para conseguir segmentaciones de calidad, detectando y separando los objetos móviles de los objetos estáticos en cada una de las imágenes de las secuencias de vídeo, en la literatura se utilizan distintas estrategias que hacen uso, no sólo de la información contenida en cada una de esas imágenes, sino también en las imágenes correspondientes a otros instantes temporales. Estas estrategias pueden clasificarse en dos grandes grupos (Wang et al., 2003), las basadas en el análisis del flujo óptico y las basadas en técnicas de substracción de fondos.

2.2.2.1. Segmentación de objetos con métodos basados en el análisis del flujo óptico

Una de las alternativas más utilizadas para llevar a cabo tareas de segmentación de objetos móviles en secuencias de vídeo es la basada en el cálculo de la velocidad local de la imagen o, lo que es lo mismo, su flujo óptico. El trabajo que sirve de partida para la mayor parte de los estudios de investigación que se han desarrollado siguiendo este tipo de análisis es el propuesto en (Horn y Schunck, 1981), existiendo a día de hoy un amplio número de propuestas, recientemente realizadas, que tratan de mejorar los resultados de estrategias anteriores (Zhou y Zhang, 2006) (Royden y Connors, 2010) (de la Escalera y Armingol, 2010) (Doshi y Bors, 2010).

El flujo óptico es una característica que describe el movimiento aparente de objetos, superficies o bordes que aparecen a lo largo de una secuencia de vídeo. Esta información es el resultado del movimiento relativo entre un observador (en nuestro caso, una cámara de vídeo) y cualquier objeto móvil o estático que forme parte de la escena filmada (Aires et al., 2008), y se puede obtener a partir del análisis de distintos tipos de información como, por ejemplo, la intensidad de los píxeles, su color o su valor de gradiente.

En la figura 2.5 se muestran tres ejemplos de campos de vectores de flujo óptico, obtenidos en el análisis del movimiento de tres secuencias de vídeo. Los dos primeros,

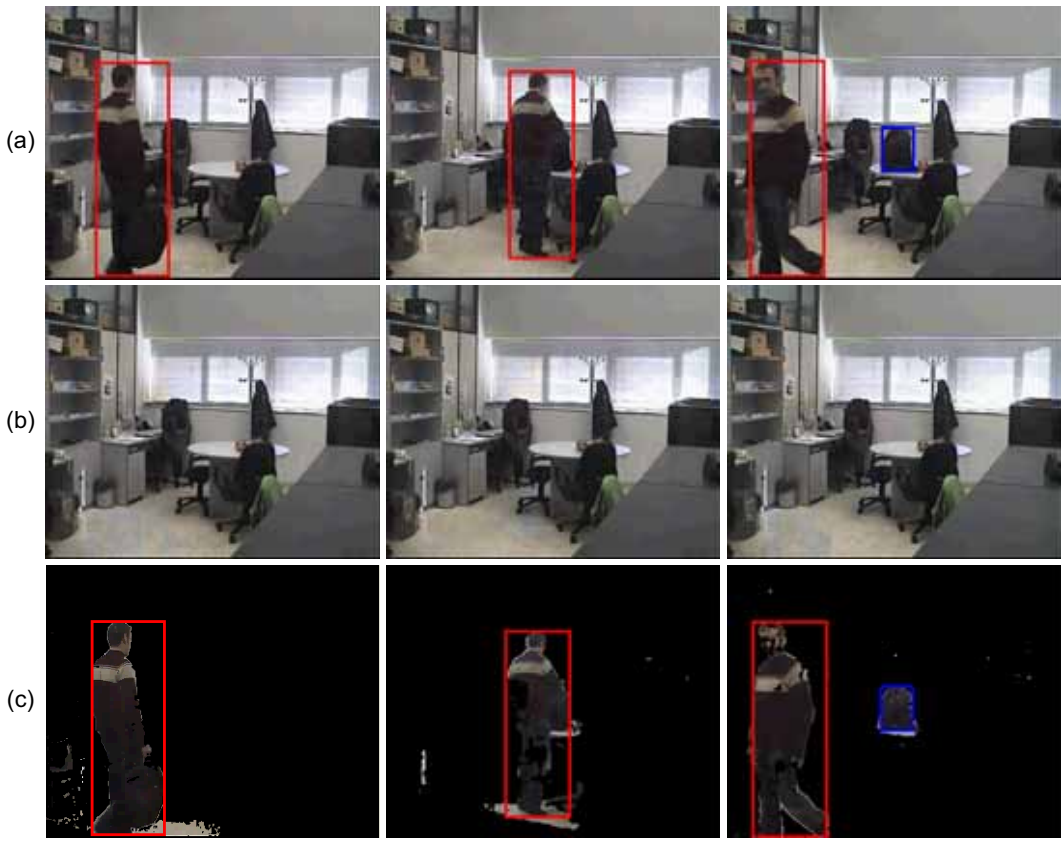


Figura 2.6: Ejemplo de segmentación de objetos utilizando técnicas de substracción de fondos. (a) Imágenes originales. (b) Fondos extraídos. (c) Objetos segmentados.

figura 2.5.a y figura 2.5.b, se corresponden con campos de vectores resultantes del análisis de puntos singulares previamente identificados en las imágenes. Sin embargo, el representado en la figura 2.5.c muestra un campo de vectores denso, en el que los vectores de flujo óptico se encuentran repartidos uniformemente por toda la imagen.

Aunque las estrategias de segmentación basadas en el análisis del flujo óptico son capaces de obtener resultados satisfactorios en un amplio abanico de situaciones, estas técnicas tienen asociada una alta complejidad y resultan demasiado sensibles en los análisis de imágenes ruidosas (Elhabian et al., 2008).

2.2.2.2. Segmentación de objetos con métodos basados en la substracción de fondos

Las técnicas de substracción de fondos tal vez sean las más utilizadas para llevar a cabo la segmentación o detección de objetos móviles en secuencias de vídeo, siendo capaces de obtener resultados de gran calidad en situaciones complejas: fondos muy dinámicos, cambios de iluminación, secuencias ruidosas, etc.

El objetivo de estas técnicas es extraer los fondos presentes en las secuencias analizadas para, de este modo, poder distinguir entre los objetos móviles y los objetos estáticos. Actualmente, en la literatura, es posible encontrar un gran número de estrategias de detección basadas en técnicas de substracción de fondos, además de trabajos que evalúan y comparan la calidad de las distintas propuestas (Piccardi, 2005) (Elhabian et al., 2008) (Bouwmans et al., 2008) (Benezeth et al., 2009) (Cristani et al., 2010), la cual suele depender de las características de las secuencias bajo análisis (la cantidad de movimiento en el fondo, los cambios de iluminación, etc.). Para determinar la calidad de estas estrategias normalmente se evalúan las siguientes características (Piccardi, 2005):

- Su velocidad de procesamiento o, lo que es lo mismo, su coste computacional asociado.
- Sus requisitos de memoria.
- La calidad de los resultados que proporcionan.

En la figura 2.6 se pueden ver los resultados obtenidos tras la aplicación de una estrategia de detección basada en la substracción de fondos sobre tres imágenes de una misma secuencia. La primera fila de imágenes de esta figura muestra las imágenes originales, en las que se puede ver cómo un objeto móvil (rodeado por un rectángulo rojo) abandona un objeto (rodeado por un círculo azul). La segunda fila de imágenes contiene los fondos obtenidos mediante el proceso de detección aplicado. En último lugar, en la tercera fila de imágenes, se ha representado el contenido móvil de la secuencia, obtenido como la diferencia entre las imágenes originales y los fondos extraídos de estas. Observando los resultados mostrados en este ejemplo se puede apreciar que tanto el objeto móvil como el objeto abandonado son correctamente detectados. Además, debido a las características de la estrategia de substracción que se ha utilizado (Stauffer y Grimson, 2002a), en los resultados correspondientes a la tercera de las imágenes mostradas se puede observar que el objeto abandonado sigue siendo clasificado como parte del contenido móvil de la imagen, ya que el fondo requiere de un tiempo mínimo de actualización. Por último, prestando atención a los detalles de los resultados mostrados en la tercera fila de imágenes, figura 2.6.c, se puede ver que los objetos móviles tienen algunos "agujeros" (píxeles móviles no detectados como tales) y que, además, también aparecen clasificados como objetos móviles algunas regiones del fondo (falsas detecciones). Esto pone de manifiesto que todavía es posible mejorar los resultados obtenidos con este tipo de estrategias.

Las estrategias propuestas para la detección de objetos móviles mediante técnicas de substracción de fondos pueden clasificarse en dos grandes grupos (Piccardi, 2005). Las que tratan de describir las variaciones de cada píxel con un único modo (técnicas unimodales), y las que lo hacen utilizando múltiples modos (técnicas multimodales).

Técnicas unimodales Son las técnicas más sencillas y, fundamentalmente, persiguen maximizar la velocidad de cálculo que requieren y reducir la memoria que necesitan. Estas técnicas son capaces de proporcionar buenos resultados en secuencias de corta duración en las que los fondos no sufren variaciones demasiado significativas (Elhabian et al., 2008). Sin embargo, los resultados que proporcionan en secuencias ruidosas, en las que aparecen cambios de iluminación, o en las que el fondo no es totalmente estático (secuencias que contienen elementos móviles en el fondo como, por ejemplo, lluvia, nieve, banderas ondeantes

o vegetación movida por el viento), son de calidad insuficiente (Parks y Fels, 2008). Dentro del conjunto de técnicas unimodales pueden destacarse dos grupos de estrategias (Cristani et al., 2010):

- ***Running Gaussian Average*** (Tang et al., 2007): Tratan de ajustar cada uno de los píxeles de la imagen a una distribución gaussiana para, de este modo, estimar la función densidad de probabilidad del fondo. Entre las ventajas de este método de segmentación se encuentran su gran sencillez de implementación, su rapidez y su bajo coste de memoria. Sin embargo, este método no ofrece buenos resultados en situaciones complicadas, como secuencias con cambios de iluminación, o situaciones en las que se producen pequeños movimientos en el fondo.
- ***Temporal Median Filter***: Estas técnicas utilizan filtros que calculan la mediana de los valores de los píxeles a lo largo del tiempo. Por ejemplo, en (Lo y Velastin, 2002) y en (Cucchiara et al., 2003) se propone la utilización de una mediana, generada a partir de los últimos valores de cada píxel, para conformar el fondo estático de la secuencia. La principal desventaja de estos métodos es su elevado coste computacional ya que, para cada píxel de la imagen actual, se deben tener almacenados muchos valores previos para poder calcular la mediana. Además, este tipo de filtrado no permite una caracterización estadística, por lo que no proporciona ninguna medida de desviación de los valores obtenidos.

Técnicas multimodales Estas técnicas, propuestas más recientemente, resuelven muchas de las limitaciones existentes en las estrategias unimodales (Benezeth et al., 2009). Se basan en el modelado de varios estados para cada píxel, haciendo posible su correcta clasificación como parte del fondo de la secuencia incluso en situaciones en las que el fondo no es estático (lluvia, vegetación movida por el viento, cambios de iluminación, etc.).

Dentro de estas estrategias de detección también es posible diferenciar entre dos grupos de técnicas (Piccardi, 2005): las que tratan de estimar la función densidad de probabilidad del fondo con modelos paramétricos, y las que lo hacen sin asumir ningún modelo en concreto.

- **Métodos basados en el modelado paramétrico**: Tratan de modelar las variaciones de cada píxel de la imagen asumiendo que estas variaciones siguen una densidad de probabilidad multimodal concreta. De esta forma, son capaces de proporcionar resultados de calidad en situaciones en las que el fondo de las secuencias contiene zonas dinámicas o en las que se producen cambios de iluminación. Algunas de las técnicas de modelado paramétrico más populares son los Modelos Ocultos de Markov (*Hidden Markov Models*, *HMM*) y la Mezcla de Gaussianas (*Mixture of Gaussians*, *MoG*).
  - *Mezcla de Gaussianas* (Bouttefroy et al., 2010) (Greggio et al., 2010): Fue propuesta por primera vez en (Grimson y Stauffer, 1999) y ha sido una de las más utilizadas como base para las nuevas estrategias de segmentación propuestas durante los últimos años. Esta estrategia hace uso de una combinación de gaussianas para tratar de representar, de forma adaptativa, las variaciones de los píxeles que contienen partes del fondo que no son estáticas. De este modo, es capaz de mejorar los resultados en situaciones en las que las técnicas unimodales ofrecen resultados

de baja calidad.
Sin embargo, también tiene algunas limitaciones. La mezcla de gaussianas no es lo suficientemente flexible como para modelar todo tipo de variaciones en el fondo ya que, en algunos casos, requiere un número demasiado alto de gaussianas para conseguirlo. Además, estos métodos tienen asociado un alto coste de computación y de memoria, debido a que cada píxel tiene asociadas varias gaussianas, cada una de ellas regida por múltiples parámetros, que deben ser actualizadas y comparadas entre sí en el análisis de cada nueva imagen.

- *Modelos Ocultos de Markov* (Bicego et al., 2006) (Utasi y Czuni, 2009) (Nascimento et al., 2010): Al igual que los métodos basados en la mezcla de gaussianas, tratan de representar las variaciones de los fondos no estáticos. En este caso, lo hacen representando los cambios posibles de cada píxel en forma de estados: día o noche, sol o lluvia, etc. De nuevo, estos métodos tienen limitaciones que han de ser tenidas en consideración: tanto la selección de un modelo adecuado como el proceso de inicialización que requieren son de gran complejidad y, además, incluyen una etapa de aprendizaje excesivamente lenta.

- **Métodos basados en el modelado no paramétrico**: Como alternativa a los métodos de modelado paramétrico, en los que se trata de modelar las variaciones de los píxeles mediante una distribución estadística concreta, existen los métodos de modelado no paramétrico. Estos métodos estiman la función densidad de probabilidad directamente, a partir de los datos, sin aplicar ninguna información a priori sobre cuál puede ser la distribución resultante. De este modo se evita tener que elegir un modelo adecuado y la selección de unos parámetros para el mismo (Elgammal et al., 2002).
Haciendo uso de estas técnicas de modelado se consigue mejorar la calidad de los resultados en situaciones en las que el modelado paramétrico no es capaz de adaptarse lo suficiente a las variaciones reales de los píxeles. Por este motivo, a lo largo de los últimos años se han publicado numerosos trabajos proponiendo estrategias de detección de objetos móviles basadas en técnicas de segmentación no paramétricas (Elgammal et al., 2002) (Mittal y Paragios, 2004) (Martel-Brisson y Zaccarin, 2008) (Tavakkoli et al., 2009).
Por otro lado, se debe tener en cuenta que estas técnicas también tienen algunos inconvenientes. Para cada píxel en cada imagen, es necesario calcular la media de un gran número de *kernels* centrados en cada muestra almacenada, resultando esto en un elevado coste de computación y de memoria (Elgammal et al., 2002).
Además, se necesita especificar un tamaño para la ventana temporal de muestras: con ventanas de mayor tamaño se mejoran los resultados, pero se incrementa el coste de computación y la necesidad de memoria. A estos inconvenientes hay que añadir el hecho de que las dependencias espaciales de los píxeles no suelen ser explotadas, y que la presencia de reflejos y sombras provocados por los objetos móviles en sus desplazamientos empeoran la calidad de los resultados.

2.3. Conclusiones

En este capítulo se ha presentado el estado del arte asociado a las estrategias de segmentación temporal de secuencias y de detección de objetos móviles desarrolladas a lo largo de la presente tesis.

En cuanto al análisis de las estrategias de segmentación temporal de secuencias de vídeo, se ha prestado especial atención a los trabajos orientados a la detección de las transiciones que hacen de frontera entre las tomas de las que constan las secuencias. De dicho análisis se puede extraer que, aunque existe una gran cantidad de estrategias de detección de transiciones, tanto abruptas como graduales, que proporcionan resultados de gran calidad (detectando la mayor parte de las transiciones y evitando la obtención de falsas detecciones), únicamente detectan ciertos tipos de transiciones en ciertos tipos de secuencias. Además, estas estrategias requieren utilizar métodos complejos o algoritmos basados en la combinación de múltiples técnicas, lo cual hace que su coste computacional sea excesivamente elevado. Por lo tanto, dada la creciente demanda de aplicaciones que requieren trabajar con grandes colecciones de vídeo, es necesario el desarrollo de nuevas estrategias de segmentación temporal que sean capaces de proporcionar resultados de la mejor calidad posible y a gran velocidad.

En relación con estado del arte correspondiente a las estrategias de segmentación de objetos en imágenes, se ha prestado especial interés a las estrategias de detección de objetos móviles en secuencias de vídeo. Como resultado de este análisis se puede concluir que, aunque existen técnicas capaces de obtener resultados satisfactorios en distintos escenarios y en presencia de fondos multimodales, las estrategias propuestas presentan numerosas limitaciones que han de ser tenidas en cuenta: los métodos paramétricos no son capaces de modelar correctamente muchas de las variaciones multimodales del fondo, constan de complejas etapas de inicialización y su calidad depende de los valores asignados a los numerosos parámetros de los que hacen uso; los métodos no paramétricos mejoran la calidad de las detecciones en presencia de fondos no estáticos, pero tienen asociados unos costes de memoria y computación excesivamente elevados y, además, dan lugar a un elevado número de falsas detecciones debidas a las sombras y a los reflejos provocados por los objetos móviles. Sin embargo, dado que la detección de objetos móviles es una etapa clave en muchas de las aplicaciones basadas en el procesamiento de imágenes y teniendo en cuenta la gran demanda de dichas aplicaciones en los últimos años, se requieren nuevas estrategias de detección que permitan obtener resultados de calidad, a gran velocidad y en el mayor número posible de escenarios, incluso en presencia de fondos muy dinámicos, cambios de iluminación, sombras, u objetos móviles con gran parecido a las regiones del fondo sobre las que se sitúan.

Capítulo 3

Segmentación temporal de secuencias de vídeo

El tiempo saca a luz todo lo que está oculto y encubre y esconde lo que ahora brilla con el más grande esplendor.
Quinto Horacio Flaco (65 AC-8 AC),
poeta latino.

Resumen: Actualmente existe un gran número de aplicaciones que, al trabajar con vídeos de gran tamaño, precisan tenerlos estructurados para, de ese modo, poder realizar operaciones de indexación y de búsqueda de contenidos. Por eso se ha desarrollado una estrategia de segmentación temporal de secuencias de vídeo en tomas, que se describe a lo largo del presente capítulo. Para detectar las transiciones abruptas entre tomas se lleva a cabo un análisis de las diferencias entre imágenes a nivel de píxel. En paralelo, se aplica un algoritmo que analiza la cantidad de puntos de borde significativos, permitiendo la detección de las transiciones graduales. Estos dos análisis se refuerzan con una segunda etapa basada en el análisis del movimiento entre pares de imágenes, la cual mejora la calidad de los resultados y reduce el problema de la selección de umbrales, a la vez que mantiene los requisitos computacionales que hacen que el sistema sea idóneo para su utilización en aplicaciones que requieren trabajar en tiempo real.

3.1. Introducción

A día de hoy, gracias a las mejoras informáticas y al crecimiento de Internet, han aparecido numerosas aplicaciones que trabajan con vídeos de gran tamaño (Chaisorn et al., 2009) como, por ejemplo: las librerías digitales, la enseñanza a distancia o los sistemas de información multimedia. Es por eso que, para gestionar dichos vídeos y gracias a las investigaciones llevadas a cabo durante los últimos años (Lew et al., 2006), se han desarrollado herramientas capaces de realizar operaciones de indexación y búsqueda de contenidos dentro de los mismos.

El primer paso en el proceso de indexación de un vídeo consiste en dividirlo en un conjunto de segmentos temporales (tomas), que serán utilizadas como elementos básicos para la generación del índice. Para poder realizar esta división es necesario localizar los cambios de cámara que definen las tomas, los cuáles pueden ser abruptos o graduales (Chasanis et al., 2009).

A lo largo de los últimos años han aparecido numerosas propuestas que describen algoritmos para la segmentación temporal de secuencias de vídeo en tomas. Entre estos, se pueden encontrar algunos trabajos que analizan y comparan las estrategias más relevantes del estado del arte (Koprinska y Carrato, 2001) (Lefèvre et al., 2003) (Cucchiara et al., 2004) (Snoek y Worring, 2005) (Brezeale y Cook, 2008) (Smeaton et al., 2010). Las estrategias propuestas en estos trabajos se pueden clasificar en seis grupos (Lefèvre et al., 2003): las basadas en el análisis a nivel de píxel, los métodos de análisis de histogramas, las que analizan las imágenes a nivel de bloques, los métodos basados en el análisis de características, los que analizan el movimiento entre imágenes, y los que son combinaciones de los anteriores.

Los métodos que realizan análisis a nivel de píxeles o a nivel de bloques, y los basados en histogramas, son rápidos y sencillos de implementar pero tienen algunos inconvenientes que deben ser tenidos en cuenta (Lawrence et al., 2004) como, por ejemplo, la selección de umbrales adecuados en función de las características del vídeo analizado. Si la cantidad de movimiento en el vídeo es elevada se necesitan umbrales altos, mientras que si apenas hay movimiento los umbrales deben ser bajos. Por lo tanto, es necesario seleccionar distintos valores en función de las características de cada secuencia e, incluso, de la cantidad de movimiento en las distintas partes de una misma secuencia. Si el umbral elegido es demasiado bajo, se obtendrá un número elevado de falsas detecciones. Por el contrario, si el valor utilizado es muy alto, un gran número de transiciones serán pasadas por alto (Brezeale y Cook, 2008).

Por otro lado, los métodos basados en el análisis del movimiento y los que combinan varias estrategias son capaces de obtener mejores resultados, detectando la mayor parte de las transiciones existentes y reduciendo el número de falsas detecciones. Sin embargo, estos métodos son mucho más complejos y considerablemente más lentos que los anteriores (Koprinska y Carrato, 2001). Además, estos métodos también suelen necesitar umbrales, por lo que el problema de la selección de valores adecuados para estos umbrales sigue estando presente.

El sistema de segmentación temporal presentado en este capítulo es una alternativa a estos métodos, siendo capaz de detectar correctamente la mayor parte de las transiciones, con una baja dependencia de los valores asignados a los umbrales utilizados, y con un coste computacional que permite su utilización en aplicaciones que requieren operar a gran velocidad. En una primera etapa se aplica un análisis de diferencias, a nivel de píxel, entre imágenes consecutivas. Este análisis permite la detección de las transiciones abruptas entre tomas. En paralelo se realiza un análisis de la cantidad de puntos de borde significativos en las imágenes, el cual permite localizar las transiciones graduales. Ambos análisis son reforzados con una segunda etapa basada en un análisis del movimiento, la cual es aplicada únicamente sobre las transiciones previamente detectadas. Esta segunda etapa reduce el

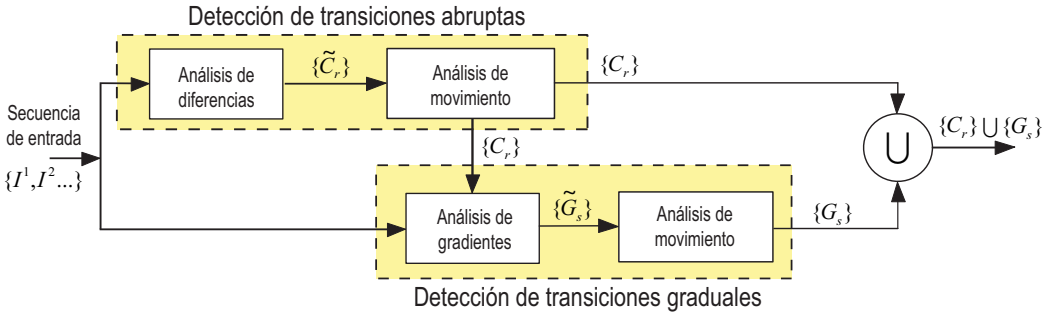


Figura 3.1: Arquitectura del sistema para la segmentación temporal de secuencias de vídeo en tomas.

problema de la selección de umbrales sin suponer un gran aumento de la carga computacional final, ya que el análisis de movimiento, que a priori puede parecer costoso, se aplica únicamente sobre algunas imágenes preseleccionadas.

Para probar la calidad y la eficiencia computacional de la estrategia propuesta se ha utilizado una base de datos compuesta por más de 30 secuencias. Dichas secuencias tienen una duración total de aproximadamente 3 horas, contienen más de 1000 tomas y poseen ciertas características que dificultan el correcto funcionamiento de los algoritmos clásicos de detección de transiciones entre tomas como, por ejemplo, objetos móviles de gran tamaño y movimientos rápidos de la cámara.

El capítulo está organizado del siguiente modo: en la sección 3.2 se describe la arquitectura del sistema propuesto; la sección 3.3 detalla la estrategia para la detección de transiciones abruptas a partir del análisis de diferencias entre imágenes consecutivas; en la sección 3.4 se describe el análisis de gradientes que permite identificar las transiciones graduales; a continuación, en la sección 3.5 se analiza la problemática en la selección de umbrales y el modo en el que la estrategia presentada la resuelve; la descripción correspondiente al análisis del movimiento entre imágenes se describe en la sección 3.6; por último, en las secciones 3.7 y 3.8 se presentan, respectivamente, los resultados obtenidos y las conclusiones.

3.2. Descripción del sistema

El sistema desarrollado combina dos estrategias que hacen uso de técnicas de bajo nivel, con una etapa basada en el análisis del movimiento entre pares de imágenes preseleccionadas. De este modo, es capaz de detectar eficientemente las transiciones, abruptas y graduales, que separan las tomas que constituyen una secuencia de vídeo. En la figura 3.1 se muestra un diagrama de bloques con una descripción detallada de los distintos módulos de los que consta dicho sistema. La entrada al sistema son las imágenes de la secuencia bajo análisis, $\{I^1, I^2 \ldots\}$, las cuales son introducidas en dos líneas de análisis en paralelo. La representada en la parte superior de la figura contiene las etapas utilizadas para la detección de las transiciones abruptas, mientras que la representada en la parte inferior muestra las etapas correspondientes a la detección de las transiciones graduales.

Para localizar las transiciones abruptas existentes, en primer lugar, se analizan las diferencias de intensidad entre imágenes consecutivas a nivel de píxel. Como resultado de este análisis se obtienen pares de imágenes, $\{\tilde{C}_r(I^{r_1}, I^{r_2})\}$, candidatas a delimitar una transición abrupta, donde r es el identificador de cada par, y en los que las imágenes I^{r_1} e I^{r_2}, con $r_2 = r_1 + 1$, son los límites de la transición candidata. En un segundo paso, aplicando una etapa de análisis del movimiento sobre cada par de imágenes candidatas, se decide si la transición realmente existe, o si es una falsa detección. Si tras este análisis se determina que la transición es correcta, se añade al conjunto de transiciones abruptas finales, $\{C_r(I^{r_1}, I^{r_2})\}$.

En el caso de las transiciones graduales, el primer paso consiste en extraer la información de gradientes de cada imagen de la secuencia. El análisis de esta información da comienzo cada vez que una transición abrupta es identificada, y se lleva cabo entre la última imagen cuyos gradientes fueron analizados y la transición abrupta recién detectada. Con este análisis se obtienen pares de imágenes que son candidatas a delimitar transiciones graduales de la secuencia, $\{\tilde{G}_s(I^{s_1}, I^{s_2})\}$, donde s es el identificador de cada transición gradual candidata. En este caso, I^{s_1} e I^{s_2}, con $s_2 = s_1 + N_g - 1$, son las imágenes que dan comienzo y fin, respectivamente, a cada transición gradual, siendo N_g el número de imágenes de que consta la transición. Para identificar las falsas detecciones y separarlas de las correctas se aplica una etapa de análisis del movimiento, similar a la utilizada en la detección de transiciones abruptas, obteniéndose el conjunto final de transiciones graduales, $\{G_s(I^{s_1}, I^{s_2})\}$. El resultado del sistema es la unión de los dos conjuntos de transiciones finales, $\{C_r\} \cup \{G_s\}$.

Es necesario considerar que, aunque el sistema propuesto es capaz de trabajar a gran velocidad, existe un tiempo de latencia al comienzo de la etapa correspondiente al análisis de los gradientes, ya que se debe haber detectado una transición abrupta para que comience este análisis. Teniendo en cuenta que esta latencia depende de la diferencia máxima entre transiciones abruptas consecutivas, es posible asegurar un tiempo máximo de latencia mediante la inserción de transiciones abruptas artificiales.

3.3. Análisis de diferencias a nivel de píxel

Como primer paso dentro de la estrategia propuesta se aplica un algoritmo rápido y eficiente que permite localizar las posibles transiciones abruptas de la secuencia bajo análisis. La idea principal en la que se basa este algoritmo, similar a la utilizada por muchos otros algoritmos propuestos en la literatura (Brezeale y Cook, 2008), es que las imágenes que pertenecen a una misma toma son más parecidas entre sí que las imágenes pertenecientes a tomas distintas. Por lo tanto, calculando las diferencias de intensidad entre píxeles situados en la misma posición espacial de imágenes consecutivas, es posible determinar cuándo un número suficientemente grande de píxeles ha cambiado significativamente, siendo esto una clara evidencia de la presencia de una transición abrupta. Sin embargo, la evaluación de diferencias entre imágenes consecutivas tiene algunos problemas. Uno de ellos es la aparición de falsas detecciones debidas a la presencia de cambios de iluminación. En una situación de este tipo, de forma similar a lo que ocurre en un cambio de toma abrupto, una gran cantidad de píxeles puede sufrir variaciones notables de intensidad entre imágenes consecutivas, lo que da lugar a la obtención de falsas transiciones.

Como solución a este problema se propone la utilización de una métrica que, comparando las variaciones de intensidad entre imágenes consecutivas con respecto a los valores medios de intensidad de las imágenes, resulta invariante a los cambios de iluminación. Dicha métrica se define como,

$$M_p(I^n|I^{n-1}) = \frac{1}{HW}\sum_{h,w}\rho^n_{h,w}, \tag{3.1}$$

donde I^n e I^{n-1} son las imágenes consecutivas a comparar, H y W son el alto y el ancho de las imágenes, el par (h, w) representa las coordenadas espaciales de cada píxel, y ρ es un parámetro que puede valer 1, -1 ó 0 en función de las diferencias entre los valores de intensidad de los píxeles y los valores medios de intensidad de cada imagen:

$$\rho^n_{h,w} = \begin{cases} 1 & \text{si } \operatorname{sign}(I^n_{h,w}-\mu^n) = \operatorname{sign}(I^{n-1}_{h,w}-\mu^{n-1}) \wedge \left|I^n_{h,w}-\mu^n\right| > T_{n1} \\ -1 & \text{si } \operatorname{sign}(I^n_{h,w}-\mu^n) \neq \operatorname{sign}(I^{n-1}_{h,w}-\mu^{n-1}) \wedge \left|I^n_{h,w}-\mu^n\right| > T_{n1} \\ 0 & \text{otros casos} \end{cases} \tag{3.2}$$

donde μ^n es el valor medio de intensidad de la imagen I^n, y T_{n1} es un umbral de ruido, con un valor típico inferior a 3, que evita tener en cuenta las pequeñas variaciones de intensidad.

En imágenes consecutivas pertenecientes a una misma toma, la relación entre los valores de intensidad de los píxeles y el valor medio de intensidad de cada imagen se mantiene aproximadamente constante en cada posición espacial. Por lo tanto, en estos casos, el resultado de aplicar la expresión 3.1 son valores de M_p próximos a 1. Sin embargo, entre imágenes pertenecientes a tomas distintas, la relación entre los valores de intensidad de la mayor parte de los píxeles con respecto a los valores medios de intensidad no se mantiene, dando lugar a valores de M_p muy inferiores a 1. Es por esto que, mediante la aplicación de esta métrica, es posible identificar la presencia de cambios de toma abruptos. Por otro lado, en presencia de un cambio de iluminación, todos los píxeles afectados modifican sus valores de intensidad de forma parecida, manteniendo su relación con el valor medio de intensidad de las imágenes. Por lo tanto, utilizando la métrica descrita, las variaciones debidas a cambios de iluminación son tratadas de forma correcta, dando lugar a valores de M_p próximos a 1, del mismo modo que ocurre en el análisis de imágenes consecutivas pertenecientes a una misma toma.

A continuación se resumen los pasos a seguir para la aplicación de esta etapa del sistema:

- Para cada nueva imagen de entrada, I^n, aplicar la métrica descrita en la ecuación 3.1.
- Comparando los valores obtenidos entre las imágenes I^{n-N_w} e I^n, localizar el mínimo local, $m_{M_p}(I^C|I^{C-1})$, donde C es la posición del mínimo y $(I^C|I^{C-1})$ es el par de imágenes que lo delimitan. En este proceso, N_w determina el número de imágenes que están siendo comparadas.
- El par de imágenes correspondiente al mínimo detectado es añadido al conjunto $\{\tilde{C}_r\}$, con $I^{r_1} = I^{C-1}$ y $I^{r_2} = I^C$, si se verifica que:

$$\left|m_{M_p}(I^C|I^{C-1}) - \text{mín}\{M_p(I^i)/i \in [n-N_w, n] \ \wedge \ i \neq C\}\right| > T_p \tag{3.3}$$

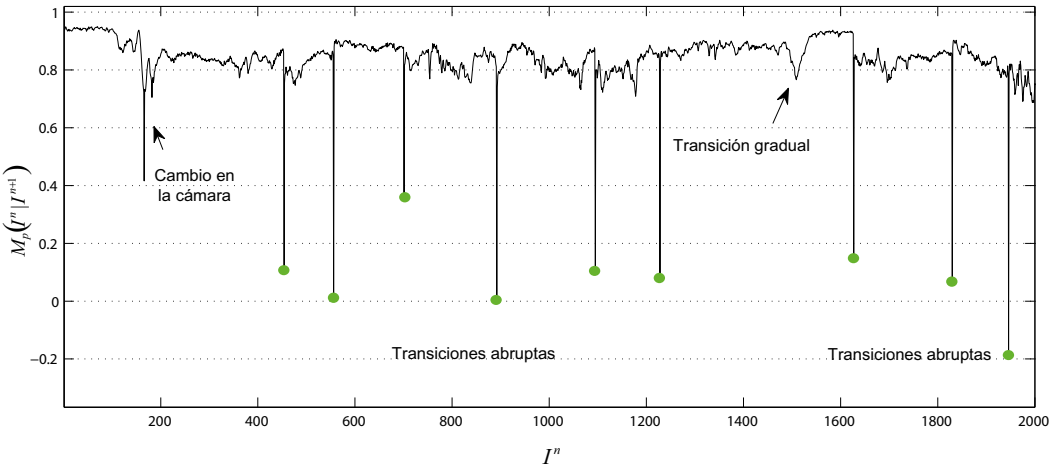


Figura 3.2: Análisis de diferencias a nivel de píxel sobre una secuencia de 2000 imágenes con 9 transiciones abruptas.

siendo T_p el umbral que determina cuándo alguno de los valores de la métrica destaca sobre los de su entorno. La influencia en los resultados de los valores asignados a los parámetros N_w y T_p se discute en el apartado 3.5.

A pesar de que los resultados obtenidos con esta métrica son muy satisfactorios, detectándose la mayor parte de las transiciones abruptas existentes, existen situaciones complicadas que pueden dar lugar a errores como, por ejemplo, las debidas a variaciones en la cámara (*pan*, *tilt*, *zoom* y desplazamientos), las transiciones graduales, o la presencia en la escena de objetos móviles de gran tamaño. En estas situaciones M_p puede mostrar variaciones significativas, similares a las que aparecen en presencia de un cambio de toma abrupto, que den lugar a una falsa detección. En la figura 3.2 se muestra un ejemplo con los resultados obtenidos, tras el cálculo de la métrica descrita, sobre una secuencia de 2000 imágenes que contiene 9 transiciones abruptas. En esta figura se observa que, normalmente, los valores de la métrica están próximos a 1, salvo para algunos pares de imágenes en los que el valor obtenido es significativamente inferior. La mayor parte de estos mínimos (marcados con círculos verdes en la figura) se debe a la existencia de transiciones abruptas. Sin embargo, algunos mínimos (señalados con flechas) son el resultado de situaciones complicadas como las que han sido descritas previamente.

Tras la aplicación de esta etapa se obtienen transiciones abruptas candidatas, que podrán ser transiciones realmente existentes o falsas detecciones debidas a cambios en la cámara, a objetos móviles de gran tamaño, o a transiciones graduales. Será en la etapa correspondiente al análisis del movimiento, descrita en el apartado 3.6, donde muchas de estas falsas transiciones sean detectadas y descartadas del conjunto final de transiciones abruptas.

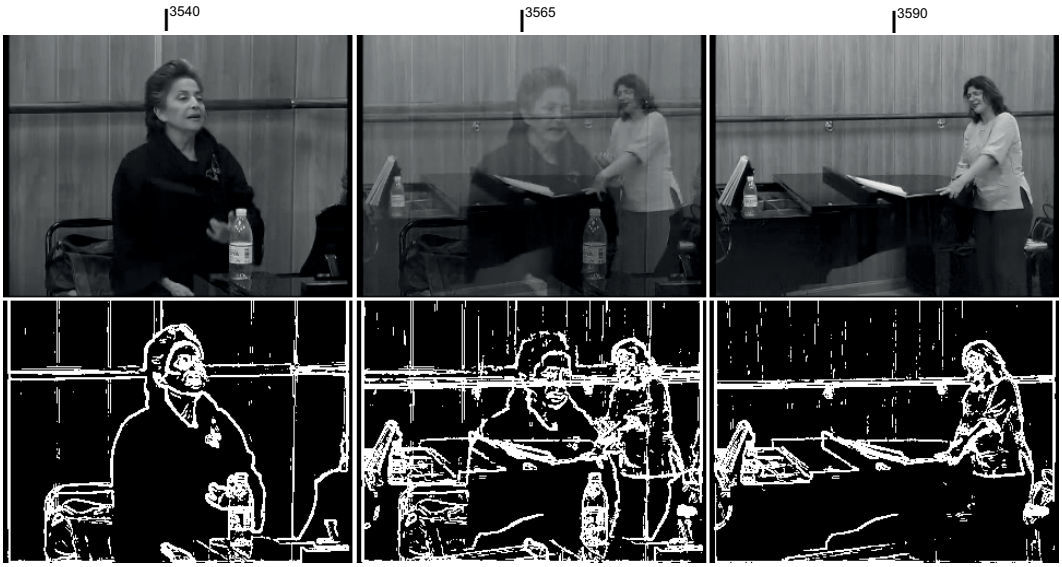


Figura 3.3: Cantidad de bordes en tres imágenes a lo largo de una transición gradual.

3.4. Análisis de gradientes

La estrategia propuesta para la detección de transiciones graduales candidatas, descrita en este apartado, se basa en el análisis de la evolución de los gradientes o bordes de las imágenes a lo largo de las secuencias de vídeo.

En una transición gradual de tipo *dissolve*, la primera toma va desapareciendo gradualmente a la vez que la segunda de las tomas aparece también de forma gradual. Partiendo de este hecho, la evolución de los bordes en una secuencia de vídeo puede proporcionar información sobre la localización de las transiciones graduales. Tras la realización de numerosos experimentos se ha comprobado que, a lo largo de una transición gradual, la cantidad de puntos de borde en las imágenes evoluciona de forma que se pasa de tener únicamente bordes pertenecientes a la primera toma (la anterior a la transición), a tener información de bordes pertenecientes únicamente a la segunda toma (la posterior a la transición), pasando por tener simultáneamente puntos de borde correspondientes a las dos tomas que separa la transición. Por lo tanto, en las imágenes donde existe información mezclada de las dos tomas aparece un máximo de puntos de gradiente. Identificando este tipo de variaciones a lo largo de una secuencia de vídeo es posible localizar las transiciones graduales que existen en la misma. En la figura 3.3 se muestra un ejemplo con tres imágenes pertenecientes a una misma transición gradual y sus correspondientes imágenes de gradientes. Prestando atención a los gradientes obtenidos de las imágenes I^{3540} e I^{3590}, límites de la transición gradual, se observa que únicamente contienen puntos de borde correspondientes a la toma anterior a la transición, en el caso de la primera, y a la toma posterior, en el caso de la segunda. Sin embargo, en el caso de la imagen I^{3565}, situada en el centro de la transición, la cantidad de puntos de borde obtenidos es muy superior, ya que contiene información tanto

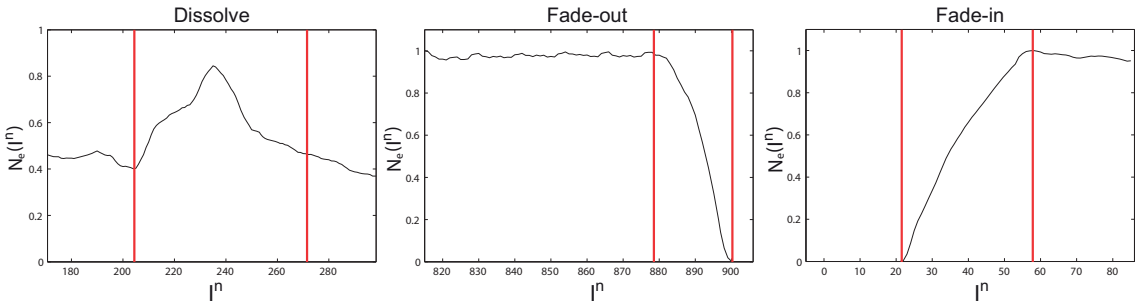


Figura 3.4: Evolución del número de puntos de borde en distintos tipos de transición gradual.

de la toma anterior como de la posterior a la transición.

En el caso de transiciones graduales de tipo *fade-in* o *fade-out* se observan situaciones similares que pueden ser identificadas de igual modo. En un *fade-in*, partiendo de cero, la cantidad de bordes aumenta progresivamente a lo largo de la transición, y en un *fade-out* la cantidad de bordes disminuye también progresivamente hasta llegar a ser nula. En la figura 3.4 se muestran 3 ejemplos de la evolución de la cantidad de puntos de borde en los 3 tipos de transición gradual mencionados: *dissolve*, *fade-in* y *fade-out*. Las líneas rojas verticales marcan el comienzo y el fin de cada transición.

En la estrategia propuesta, la información utilizada es el porcentaje de puntos de borde significativos en cada imagen de la secuencia, $N_e(I^n)$, entendiendo como significativo a todo punto que posea un gradiente con módulo superior a un umbral de ruido, T_{n2}. Tras la realización de numerosos experimentos se ha comprobado que la elección de un valor u otro para este umbral, siempre que este valor sea pequeño, no influye en los resultados obtenidos.

Los pasos a seguir para la aplicación de esta estrategia se resumen a continuación:

- Calcular el porcentaje de píxeles de la imagen, $N_e(I^n)$, con un gradiente de módulo superior a T_{n1}, entre un par de transiciones abruptas previamente detectadas.
- Aplicar un filtrado paso bajo sobre $N_e(I^n)$ para eliminar las variaciones ruidosas.
- Normalizar $N_e(I^n)$.
- Localizar el conjunto de máximos locales, $\{M_{N_e}(I^n)\}$: cada máximo es candidato a formar parte de una transición gradual.
- Para cada máximo encontrado, localizar sus mínimos anterior y posterior más cercanos, $\{M_{N_e}(I^{n-A})\}$ y $\{M_{N_e}(I^{n+B})\}$.
- Se añade una nueva transición gradual candidata al conjunto $\{\tilde{G}_s\}$, definida por $I^{s_1} = I^{n-A}$ y por $I^{s_2} = I^{n+B}$, si se cumple alguna de las dos condiciones siguientes:

$$\begin{aligned} \left|M_{N_e}(I^n) - M_{N_e}(I^{n-A})\right| > T_e \ \wedge \ A \geq N_g/2 \\ \left|M_{N_e}(I^n) - M_{N_e}(I^{n+B})\right| > T_e \ \wedge \ B \geq N_g/2 \end{aligned} \tag{3.4}$$

En estas expresiones, T_e es porcentaje mínimo que ha de variar la cantidad de puntos con gradientes significativos, en relación con el número máximo de estos puntos en

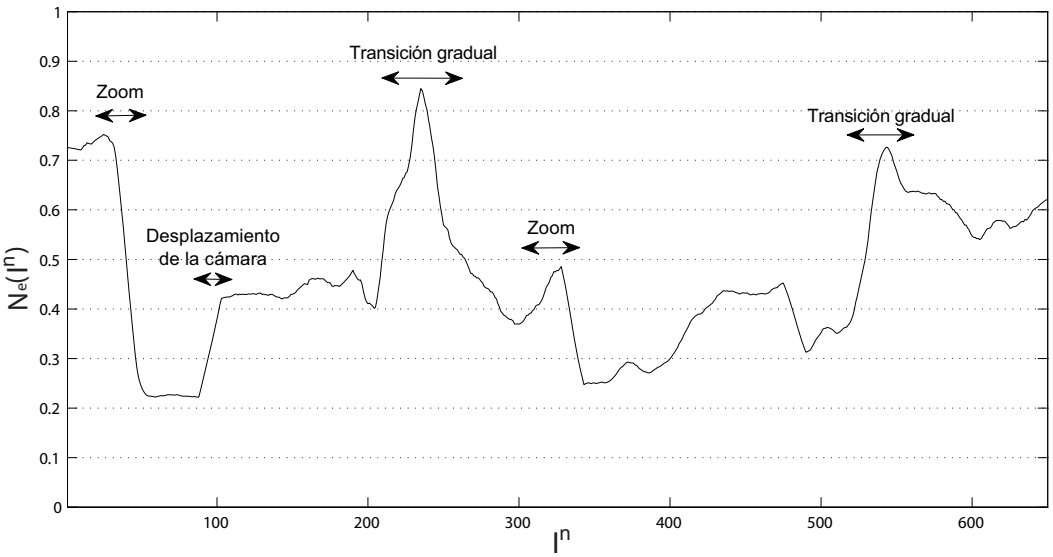


Figura 3.5: Porcentaje de puntos de borde significativos a lo largo de un tramo de vídeo comprendido entre 2 transiciones abruptas.

el intervalo analizado, para añadir una nueva transición gradual candidata. Por otro lado, el parámetro N_g determina el número mínimo de imágenes que debe tener una transición gradual para poder ser detectada. La calidad de los resultados en función de los valores asignados a estos dos parámetros se discute en el apartado 3.5.

Al realizarse este análisis sobre tramos de la secuencia que no contienen cambios abruptos de toma, se evita la aparición de variaciones bruscas en el número medio de puntos de borde, facilitándose la utilización de un único umbral, T_e, a lo largo de toda la secuencia.

Aunque, generalmente, los resultados obtenidos haciendo uso de la estrategia descrita son muy satisfactorios, la evolución de $N_e(I^n)$ puede mostrar un comportamiento similar al de las transiciones graduales en situaciones de *zoom*, movimientos *pan*/*tilt* de la cámara, o en escenas con objetos móviles de gran tamaño. Por lo tanto, estas situaciones pueden ser clasificadas de forma incorrecta como transiciones graduales. Será en la etapa posterior, basada en el análisis del movimiento, donde muchas de estas falsas detecciones serán identificadas y descartadas del conjunto final de transiciones graduales. En la figura 3.5 aparece representada la evolución de $N_e(I^n)$ a lo largo de un tramo comprendido entre 2 transiciones abruptas. Este tramo consta de 650 imágenes y contiene: 2 transiciones graduales, 2 intervalos con variaciones en el *zoom* y 1 intervalo con movimiento significativo de la cámara.

3.5. Selección de umbrales

En las técnicas de segmentación temporal, la selección de umbrales adecuados es de gran relevancia ya que, dependiendo de la cantidad de movimiento que existe en una secuencia, los

resultados obtenidos pueden ser mejores o peores. Si los valores utilizados son demasiado bajos, muchas situaciones poco deseables, como movimientos de objetos móviles o de la cámara, darán lugar a falsas detecciones. Por otro lado, si los umbrales tienen valores elevados, aumentan las posibilidades de que algunas de las transiciones existentes sean pasadas por alto.

Por ejemplo, en la figura 3.2 se muestran algunas situaciones, debidas a movimientos de la cámara y a una transición gradual, que pueden ser motivo de detección de falsas transiciones. Si el valor de T_p es demasiado bajo, estas situaciones serán clasificadas como transiciones abruptas. Sin embargo, si T_p tiene un valor demasiado elevado, algunas de las transiciones realmente existentes pueden ser ignoradas.

Además, en el caso de la estrategia propuesta para la detección de transiciones abruptas, también debe ser tenido en cuenta el parámetro N_w. Si su valor es demasiado alto, el número de transiciones abruptas no detectadas aumentará debido a que, si supera la distancia entre transiciones abruptas consecutivas, se estarán comparando valores de M_p con varios mínimos significativos. Por el contrario, si el valor de N_w es demasiado bajo se estarán comparando muy pocos pares de imágenes y, consecuentemente, aumentará el número de falsas detecciones.

En el caso de las transiciones graduales, si el valor de T_e es muy bajo habrá muchos intervalos temporales en los que las variaciones de los gradientes den lugar a falsas detecciones como, por ejemplo, los 2 *zooms* y el movimiento de la cámara que aparecen en el ejemplo de la figura 3.5. Por el contrario, si su valor es muy alto, habrá transiciones graduales que no serán detectadas.

En este caso también se debe tener en cuenta el valor de N_e. Si su valor es demasiado alto, las transiciones graduales constituidas por pocas imágenes no serán detectadas. Sin embargo, si su valor es muy bajo, el número de falsas detecciones aumentará notablemente.

Existen estrategias que, dependiendo de sus objetivos, asignan un valor u otro a los umbrales que utilizan. Con valores altos se consigue reducir las falsas detecciones, pero así también aumenta el número de transiciones no detectadas. Con valores bajos son menos los cambios de toma no detectados, pero se incrementa el número de falsas transiciones erróneamente detectadas.

En la estrategia aquí propuesta se ha decidido utilizar umbrales con valores suficientemente bajos, de forma que se asegure la correcta detección de la mayor parte de las transiciones existentes, a pesar de aumentar el número de falsas detecciones. Estas falsas detecciones serán identificadas y descartadas mediante la aplicación del módulo basado en el análisis del movimiento que será descrito en el siguiente apartado, en el que la selección de umbrales no es tan crucial como en las técnicas descritas hasta ahora. De este modo se consigue reducir el número de falsas detecciones, manteniendo la mayor parte de las transiciones existentes.

La elección de los valores más adecuados para todos estos parámetros se justifica en la sección 3.7.

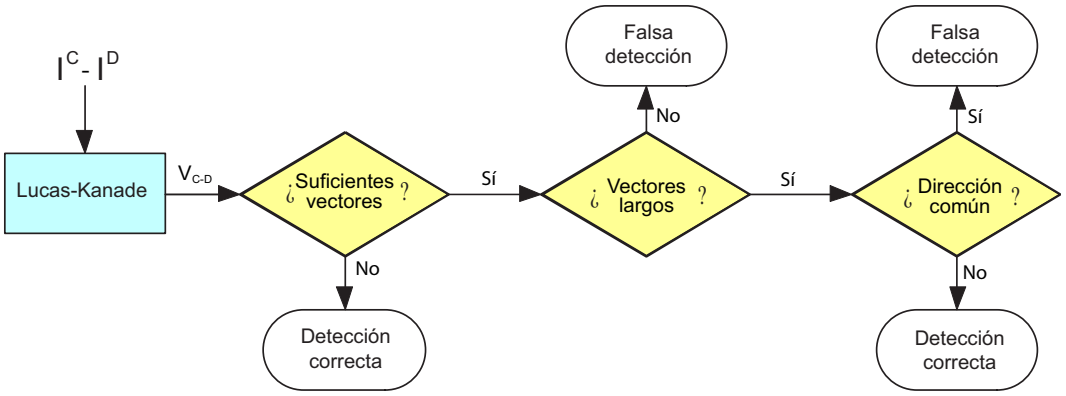


Figura 3.6: Diagrama de flujo correspondiente a la etapa de análisis del movimiento de transiciones candidatas.

3.6. Análisis del movimiento entre imágenes

Tras la aplicación de los métodos descritos en los apartados anteriores se han obtenido dos conjuntos de transiciones candidatas, uno de abruptas, $\tilde{C}_r$, y otro de graduales, $\tilde{G}_s$. Utilizando umbrales con valores suficientemente bajos se consigue que estos dos conjuntos contengan la mayor parte de las transiciones existentes. Sin embargo, en estos conjuntos también estarán incluidas algunas falsas detecciones debidas a la existencia de efectos no deseados que son consecuencia de variaciones en la cámara de adquisición o de la presencia de objetos móviles de gran tamaño en la escena. Por lo tanto, es necesario detectar y separar estas falsas detecciones de las correctas. Para ello se ha desarrollado una estrategia que permite analizar el movimiento entre pares de imágenes, la cual, al ser aplicada únicamente sobre algunas imágenes pre-seleccionadas, apenas incrementa el coste computacional del sistema.

Mediante este análisis del movimiento, tras comparar cada par de imágenes, se obtiene un campo de vectores de desplazamiento. Si este campo de vectores muestra algún tipo de coherencia, tanto espacial como de orientación, es muy probable que esté determinando algún tipo de movimiento de la cámara o de los objetos presentes en la escena. Por lo tanto, las transiciones candidatas para las que se obtenga esta coherencia serán descartadas de los conjuntos finales de transiciones, C_r y G_s.

La figura 3.6 contiene un diagrama de flujo que muestra detalladamente las etapas de las que consta la fase de análisis basada en el movimiento. Este análisis se lleva a cabo para cada par de imágenes, $(I^C - I^D)$, que limitan a las transiciones candidatas previamente detectadas. En primer lugar se aplica el algoritmo piramidal de *Lukas-Kanade* (Bouguet, 1999) sobre el par de imágenes, obteniéndose un campo de vectores de movimiento, $\{V_{C-D}(i)\}_{i=1}^{N_v}$, donde N_v es el número de vectores obtenidos. Este campo relaciona puntos característicos de la primera imagen, obtenidos mediante la aplicación del algoritmo de detección de *Harris* (Harris y Stephens, 1988), con puntos similares localizados en la segunda imagen. Para determinar si las imágenes analizadas son los límites de una transición que realmente existe

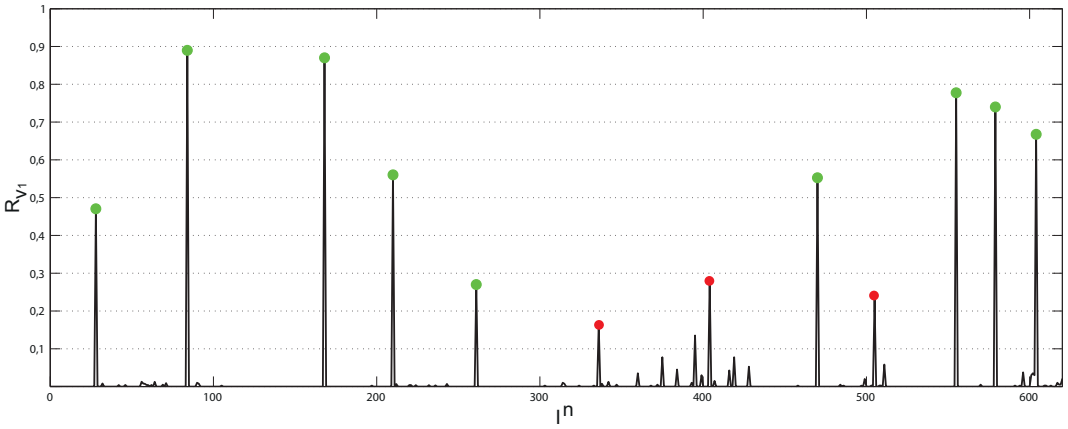


Figura 3.7: Variación de la cantidad de vectores de desplazamiento identificados a lo largo de una secuencia de vídeo.

Figura 3.8: Ejemplos de transiciones candidatas con distinto número de vectores de desplazamiento.

se analizan hasta tres características del campo de vectores de movimiento: la cantidad de vectores obtenidos, su longitud y su orientación. En las siguientes secciones se describen detalladamente las tres etapas correspondientes al análisis de estas tres características.

3.6.1. Análisis de la cantidad de vectores

En primer lugar, para cada transición candidata, se analiza el número de vectores, N_v, resultante de la comparación entre los puntos singulares de las imágenes que delimitan a dicha transición. Normalmente, en las transiciones realmente existentes, debido a la falta de semejanzas entre las imágenes comparadas, el número de vectores obtenido es muy inferior al número de características identificadas en la primera de las imágenes, I^C. Sin embargo, en las falsas transiciones, al existir un mayor número de semejanzas entre las imágenes comparadas (por pertenecer estas a una misma toma), la cantidad de vectores de desplazamiento obtenidos es mucho más parecida al número de puntos característicos localizados en

la imagen I^C. Por lo tanto, es posible determinar que una transición candidata, $I^C - I^D$, es correcta si se verifica que:

$$R_{v_1} = 1 - \frac{N_v}{N_c} > T_{v_1} \tag{3.5}$$

donde N_c es el número de puntos característicos en la imagen I^C y $T_{v_1} \in [0, 1]$ es un umbral cuyo valor se justifica en la sección 3.7.

La figura 3.7 muestra un ejemplo con los valores de R_{v_1} a lo largo de una secuencia de 630 imágenes con 9 transiciones (marcadas con puntos verdes). Este ejemplo permite comprobar que, en las comparaciones entre imágenes de distintas tomas, el valor de R_{v_1} es muy superior al del resto de la secuencia, salvo en algunos instantes en los que la falta de puntos característicos en la imagen I^C ha dado lugar a valores de R_{v_1} algo mayores (instantes señalados con puntos rojos).

En la figura 3.8 aparecen representados los vectores de desplazamiento correspondientes a la comparación de dos transiciones candidatas. De la primera comparación (figura 3.8.a), correspondiente a una transición correcta, se ha obtenido un número de vectores muy bajo en relación con el número de puntos característicos identificados en la imagen I^C. Sin embargo, de la comparación correspondiente al segundo par de imágenes (figura 3.8.b), ambas pertenecientes a misma toma, se ha obtenido un número de vectores mucho más alto.

3.6.2. Análisis de la longitud de los vectores

En el caso de obtener un valor de $R_{v_1} > T_{v_1}$, para tomar una decisión sobre la autenticidad de la transición analizada será necesario evaluar la longitud media de los vectores, S_v. Esta longitud media se obtiene aplicando la expresión:

$$S_v = \frac{1}{N_v} \sum_{i=1}^{N_v} \left(\left(\frac{L_{H_i}}{H} \right)^2 + \left(\frac{L_{W_i}}{W} \right)^2 \right)^{\frac{1}{2}} \tag{3.6}$$

en la que (L_{H_i}, L_{W_i}) son las componentes, en filas y columnas, del vector i-ésimo. Para comparar fácilmente la longitud media de los vectores en los distintos casos analizados, independientemente de que dichos análisis se hayan realizado sobre una misma secuencia o sobre secuencias con imágenes de distinta resolución espacial, las componentes de los vectores se normalizan por las dimensiones de las imágenes (H, W). Sin embargo, se ha de tener en cuenta que esta normalización hace que se dé más importancia a los vectores que describen desplazamientos en la dirección del lado de menor tamaño de las imágenes.

En las falsas detecciones debidas a situaciones en las que la escena contiene elementos móviles de gran tamaño o en las que la cámara realiza pequeños movimientos, el campo de vectores resultante del análisis de movimiento contendrá un gran número de vectores que, en relación con el tamaño de las imágenes, serán cortos. Por lo tanto, identificando esta situación es posible descartar las falsas transiciones ocasionadas por alguno de los motivos descritos.

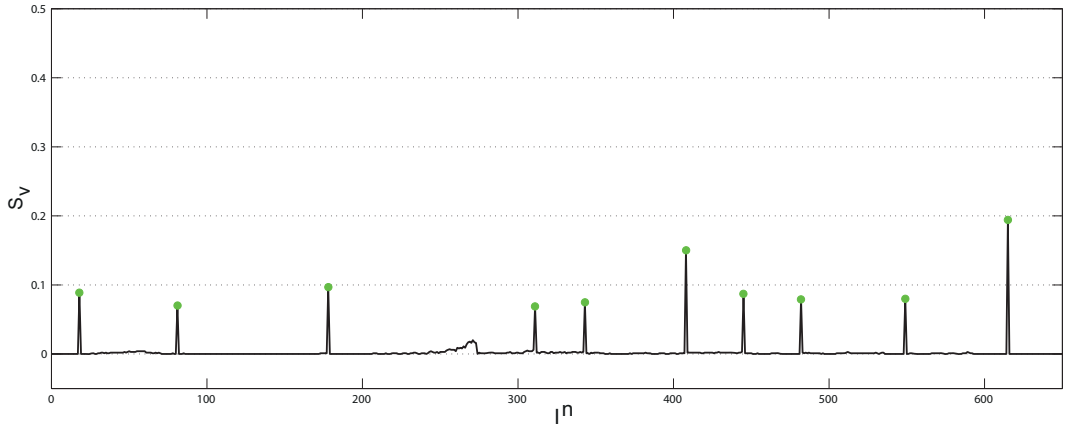


Figura 3.9: Longitud media de los vectores de desplazamiento identificados a lo largo de una secuencia de vídeo.

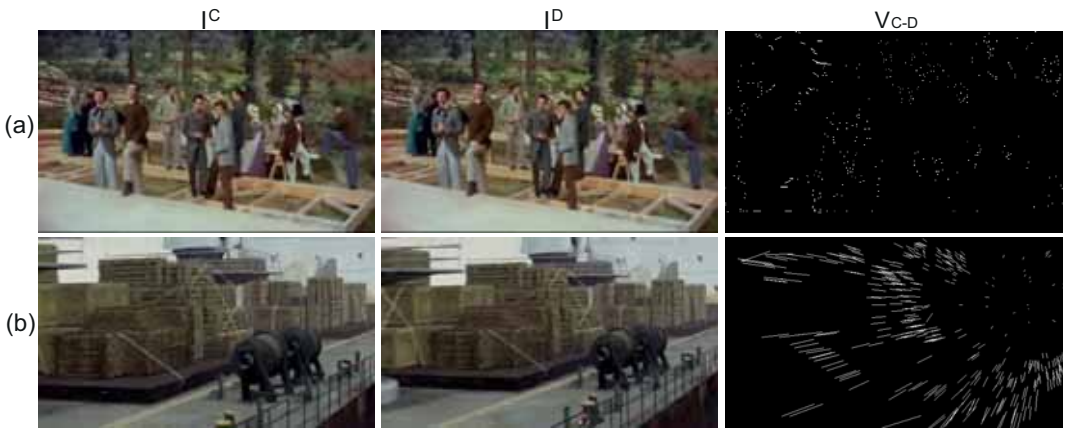


Figura 3.10: Ejemplo de campos de vectores de movimiento con numerosos vectores cortos.

La figura 3.9 muestra la longitud media de los vectores de desplazamiento, entre pares de imágenes consecutivas, a lo largo de una secuencia de 650 imágenes con 10 transiciones entre tomas (señaladas mediante puntos verdes). Este ejemplo permite apreciar que la longitud media de los vectores obtenidos de la comparación entre imágenes de distintas tomas es notablemente superior a la de los vectores obtenidos de la comparación entre imágenes de la misma toma. Por lo tanto, descartando las transiciones candidatas con una longitud media inferior a un umbral $T_{v_2} \in [0, 1]$, es posible eliminar una gran cantidad de falsas detecciones. La elección de un valor apropiado para este umbral se discute en la sección 3.7.

En la figura 3.10.a se presenta un ejemplo de falsa transición debida al movimiento de objetos móviles de grandes dimensiones. Se observa que, en esta situación, los vectores obtenidos son muy cortos. La segunda fila de imágenes de la misma figura (figura 3.10.b) muestra un resultado similar (otra falsa detección caracterizada por numerosos vectores cortos). Sin embargo, este caso, estos vectores son el resultado de un *zoom* de la cámara.

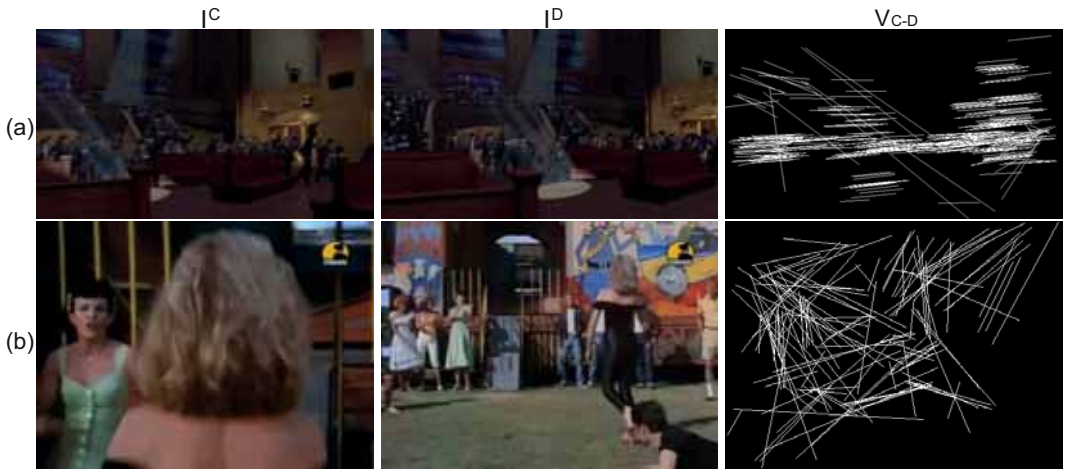


Figura 3.11: Ejemplo de campos de vectores de movimiento con numerosos vectores largos.

3.6.3. Análisis de la dirección de los vectores

Si el número de vectores es elevado ($R_{v_1} > T_{v_1}$) y, además, la mayor parte de los vectores de movimiento tiene una longitud considerable ($S_v > T_{v_2}$), para poder determinar si las imágenes comparadas delimitan o no a una transición, será necesario analizar la dirección de los vectores. Cuando la cámara realiza algún tipo de operación de tipo *pan*, *tilt* o *zoom*, o cuando ésta cambia su posición rápidamente, los vectores de movimiento entre imágenes consecutivas suelen ser largos y poseer una orientación similar. Por lo tanto, analizando la orientación del campo de vectores, es posible determinar si existe algún tipo de desplazamiento o algún cambio en la cámara que haya dado lugar a una falsa detección o, si debido a la falta de coherencia entre las imágenes, la transición analizada es correcta. Las imágenes de la figura 3.11.a muestran un ejemplo de falsa detección debida a un desplazamiento rápido de la cámara. Se puede comprobar que, con el análisis del movimiento, se han obtenido muchos vectores de gran longitud con direcciones muy parecidas. Las imágenes de la figura 3.11.b muestran otro ejemplo, correspondiente a una transición real, que ha dado lugar a un numeroso campo de vectores de movimiento que, aunque posee muchos vectores de longitud considerable, muestra orientaciones muy variadas.

Para llevar a cabo el análisis de la orientación de los vectores obtenidos, se siguen los siguientes pasos:

- Calcular las orientaciones de los vectores, $\{D_{v_i}\}_{i=1}^{N_v}$, entre 0 y 360 grados.
- Obtener la dirección dominante de los vectores, ponderando su dirección por su longitud para, de este modo, reducir la influencia de los vectores cortos:

$$\mu_v = \frac{1}{\sum_{i=1}^{N_v} (L_{H_i}^2 + L_{V_i}^2)^{\frac{1}{2}}} \sum_{i=1}^{N_v} D_{v_i} (L_{H_i}^2 + L_{V_i}^2)^{\frac{1}{2}} \tag{3.7}$$

- Obtener la desviación típica de estas orientaciones, ponderando también por la longitud

de los vectores:

$$\sigma_v = \frac{1}{\left(\sum_{i=1}^{N_v} (L_{H_i}^2 + L_{V_i}^2)^{\frac{1}{2}}\right)^2} \sum_{i=1}^{N_v} (D_{v_i} - \mu_v)^2 (L_{H_i}^2 + L_{V_i}^2)^{\frac{1}{2}} \tag{3.8}$$

- La transición bajo análisis es identificada como falsa si se cumple que $\sigma_v < T_{v_3}$, mientras que si no se cumple esta condición, se determina que la transición es correcta. La elección del valor más adecuado para el umbral $T_{v_3} \in [0^o, 360^o]$ se discute en la sección 3.7.

3.7. Resultados

Para evaluar la calidad del sistema de segmentación temporal presentado, se ha analizado una base de datos compuesta por más de 30 secuencias que contienen más de 1000 tomas. La descripción detallada de esta base de datos aparece en el apéndice A.

Como medidas de evaluación, debido a que son muy comúnmente utilizados en la literatura (Koprinska y Carrato, 2001), se han utilizado los porcentajes de *Recall* y *Precision*. Estos porcentajes se calculan haciendo uso de las siguientes expresiones:

$$\text{Recall} = 100 \frac{\text{CD}}{\text{CD} + \text{ND}} \,\% \tag{3.9}$$

$$\text{Precision} = 100 \frac{\text{CD}}{\text{CD} + \text{FD}} \,\% \tag{3.10}$$

en las que CD es el número de transiciones correctamente detectadas, ND es el número de transiciones no detectadas, y FD es el número de falsas detecciones obtenidas tras el análisis. El porcentaje de *Recall* relaciona la cantidad de transiciones correctamente detectadas con el número total de transiciones existentes. El porcentaje de *Precision* relaciona las transiciones correctamente detectadas con el número total de transiciones detectadas.

Además de estos dos, se ha utilizado un tercer porcentaje, denominado F (Teng et al., 2008), que evalúa de forma conjunta los resultados de *Recall* y de *Precision*. Se define como:

$$F = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \,\% \tag{3.11}$$

A lo largo de esta sección de resultados se analiza tanto la calidad de las detecciones obtenidas mediante la aplicación de las distintas etapas descritas en el presente capítulo, como la influencia en los resultados de los valores asignados a las variables y umbrales utilizados en dichas etapas. El análisis correspondiente a la etapa de detección de transiciones abruptas candidatas se presenta en la sección 3.7.1. A continuación, en la sección 3.7.2 se presenta el análisis correspondiente a la etapa de detección de transiciones graduales candidatas. Acto seguido, en la sección 3.7.3 se analizan las mejoras obtenidas mediante el análisis del movimiento aplicado sobre las transiciones candidatas. Por último, en la sección 3.7.4 se evalúa la eficiencia computacional de la estrategia propuesta.

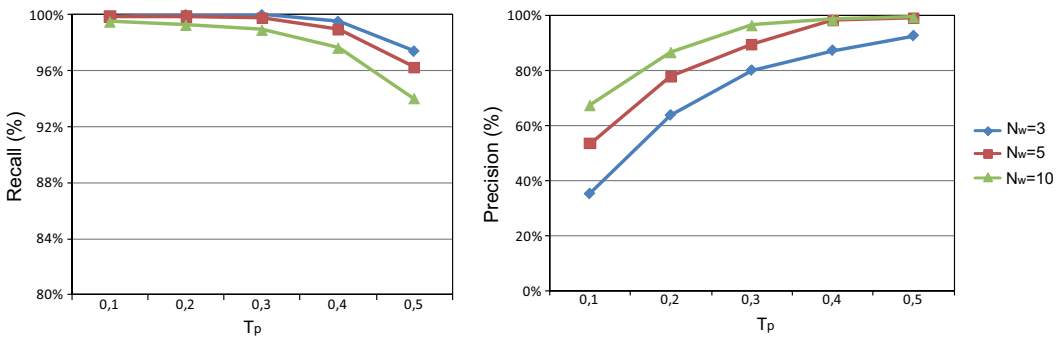


Figura 3.12: Porcentajes de acierto y precisión para distintos valores de N_w y T_p.

Tipo de secuencia	Nº de trans. abruptas	CD	FD	ND	*Recall* (%)	*Precision* (%)	F (%)
Dibujos	279	279	50	0	100	84,80	91,78
Musicales	206	204	16	2	99,03	92,73	95,77
Reportajes	82	82	7	0	100	92,13	95,91
Otros	284	284	26	0	100	91,61	95,62
Total	851	849	99	2	99,76	89,56	94,39

Tabla 3.1: Resultados correspondientes a la detección de transiciones abruptas antes del análisis de movimiento.

3.7.1. Detección de transiciones abruptas candidatas

Tal y como se ha mencionado en el apartado 3.5, la elección de unos valores adecuados para las variables N_w y T_p es fundamental. Por ese motivo se ha llevado a cabo un análisis de los resultados obtenidos para distintos valores de estas variables.

La figura 3.12 muestra dos gráficas con los valores de *Recall* y *Precision* obtenidos tras la aplicación de la primera etapa para la detección de las transiciones abruptas, utilizando distintos valores de N_w y T_p. Los resultados de estas gráficas ponen de manifiesto que el número de transiciones no detectadas aumenta a medida que aumentan los valores de N_w y de T_p. Como el objetivo de esta etapa es detectar el mayor número posible de las transiciones existentes (altos porcentajes de *Recall*), se ha decidido utilizar $N_w = 5$ y $T_p = 0,3$, ya que con valores inferiores de cualquiera de los dos parámetros no se mejora la cantidad de detecciones correctas (no aumenta el porcentaje de *Recall*) y, sin embargo, sí que se incrementa el número de falsas detecciones (se reduce el porcentaje de *Precision*). Los resultados correspondientes a la aplicación de esta primera etapa, haciendo uso de estos valores, aparecen resumidos en la tabla 3.1.

3.7.2. Detección de transiciones graduales candidatas

En la etapa para la detección de las transiciones graduales candidatas, recordando lo mencionado en la sección 3.5, los valores asignados a los parámetros N_e y T_e tienen una impor-

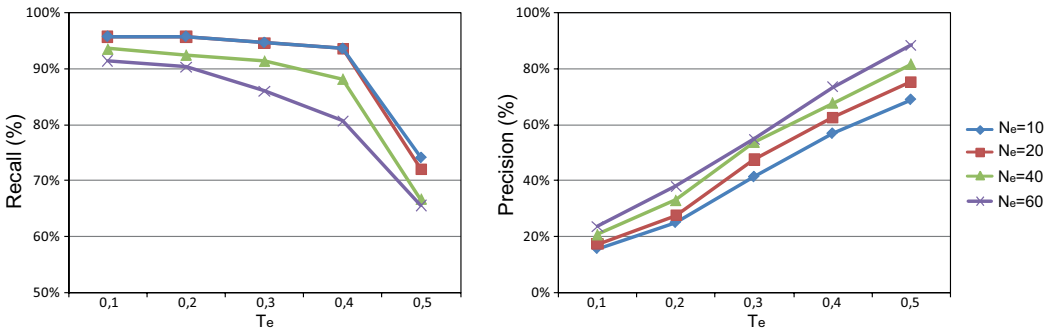


Figura 3.13: Porcentajes de acierto y precisión para distintos valores de N_e y T_e.

Tipo de secuencia	Nº de trans. graduales	CD	FD	ND	*Recall* (%)	*Precision* (%)	F (%)
Dibujos	7	6	11	1	85,71	35,29	50,00
Musicales	8	8	9	0	100	47,06	64,00
Reportajes	77	72	19	5	93,51	79,12	85,71
Otros	1	1	13	0	100	7,14	13,33
Total	93	87	52	6	93,55	62,59	75,00

Tabla 3.2: Resultados correspondientes a la detección de transiciones graduales antes del análisis de movimiento.

tante influencia en la calidad de los resultados.

En la figura 3.13 se presenta una comparación entre los resultados obtenidos tras asignarles distintos valores. Estos resultados permiten apreciar que a medida que aumenta los valores de N_e y de T_e se incrementa el número de detecciones no detectadas (menor *Recall*) y se reduce el número de falsas detecciones (mayor *Precision*). En este caso, al igual que en el analizado en la sección 3.7.1 para el caso de las transiciones abruptas, también se persigue obtener porcentajes de *Recall* lo más altos posibles. Dado que si se utilizan valores inferiores a $T_e = 0{,}4$ y $N_e = 20$, el porcentaje de *Recall* apenas mejora y, sin embargo, el de *Precision* empeora notablemente, se ha decidido utilizar estos valores. El resumen con los resultados obtenidos en esta etapa haciendo uso de estos valores se muestra en la tabla 3.2.

3.7.3. Selección de las transiciones finales

Para llevar a cabo una elección adecuada de los valores asignados a los umbrales que determinan si una transición candidata es, finalmente, clasificada como correcta o como falsa (T_{v_1}, T_{v_2} y T_{v_3}), se han analizado las características de los campos de vectores asociados a todas las transiciones candidatas obtenidas. El resumen de estas características es el presentado mediante las gráficas de la figura 3.14.

En la primera de estas gráficas se analiza la influencia del umbral T_{v_1} en los resultados. Los datos representados en esta gráfica muestran: por un lado, el porcentaje de transiciones

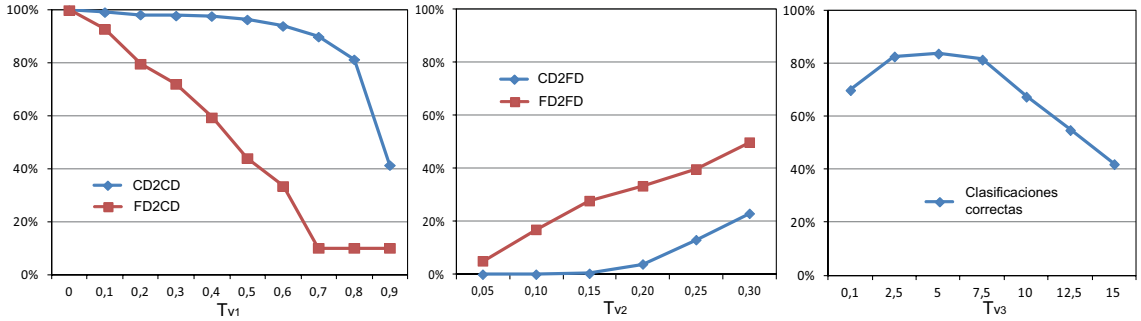


Figura 3.14: Resumen de características de los vectores de desplazamiento de las transiciones candidatas.

Tipo de secuencia	Nº de trans. abruptas	CD	FD	ND	*Recall* (%)	*Precision* (%)	F (%)
Dibujos	279	274	16	5	98,21	94,48	96,31
Musicales	206	200	4	6	97,09	98,04	97,56
Reportajes	82	76	0	6	92,68	100	96,20
Otros	284	283	0	1	99,65	100	99,82
Total	851	833	20	18	97,88	97,66	97,77

Tabla 3.3: Resultados correspondientes a la detección de transiciones abruptas después del análisis de movimiento.

Tipo de secuencia	Nº de trans. graduales	CD	FD	ND	*Recall* (%)	*Precision* (%)	F (%)
Dibujos	7	6	3	1	85,71	66,67	75
Musicales	8	8	3	0	100	72,73	84,21
Reportajes	77	71	3	6	92,21	95,95	94,04
Otros	1	1	2	0	100	33,33	50
Total	93	86	11	7	92,47	88,66	90,53

Tabla 3.4: Resultados correspondientes a la detección de transiciones graduales después del análisis de movimiento.

candidatas correctas que son clasificadas como tales (CD2CD); por otro lado, el porcentaje de falsas transiciones que, de forma errónea, son clasificadas como correctas (FD2CD). Recordando lo expuesto en la sección 3.6.2, la etapa de análisis en la que se utiliza este umbral tiene como objetivo identificar los casos en los que la relación entre puntos singulares y vectores de movimiento es lo suficientemente alta como para decidir que una transición candidata es correcta. Sin embargo, se debe tratar de minimizar el número de transiciones falsas clasificadas como correctas. Si $T_{v_1} \geq 0,7$, el porcentaje de falsas transiciones erróneamente clasificadas como correctas es aproximadamente constante (entorno al 10 %), mientras que para valores inferiores de T_{v_1} este porcentaje aumenta rápidamente. Por lo tanto, sea cual

Tipo de secuencia	Nº de trans.	CD	FD	ND	*Recall* (%)	*Precision* (%)	F (%)
Dibujos	286	280	19	6	97,90	93,65	95,73
Musicales	214	208	7	6	97,20	96,74	96,97
Reportajes	159	147	3	12	92,45	98	95,15
Otros	285	284	2	1	99,65	99,30	99,47
Total	944	919	31	25	97,35	96,74	97,04

Tabla 3.5: Resultados globales, correspondientes a la detección conjunta de transiciones abruptas y graduales.

sea el valor de T_{v_1}, al menos el 10 % de las falsas transiciones serán mal clasificadas (típicamente, transiciones en las que apenas se han detectado puntos singulares). Sin embargo, cuanto menor es el valor T_{v_1} mayor es el porcentaje de transiciones correctas etiquetadas como tales. Es por eso que hemos decidido utilizar $T_{v_1} = 0,7$.

En la segunda gráfica de la figura 3.14 se muestra, para distintos valores de T_{v_2}, el porcentaje de transiciones candidatas que, siendo correctas, son clasificadas como falsas (CD2FD), y las transiciones candidatas falsas que son etiquetadas como tales (FD2FD). Los datos obtenidos muestran que para valores de T_{v_2} inferiores a $0,15$ el número de transiciones correctas clasificadas como falsas es prácticamente nulo. Sin embargo, cuanto mayor es T_{v_2} mayor es el número de falsas detecciones descartadas. Por lo tanto, para descartar el máximo número posible de falsas detecciones sin apenas incrementar el de transiciones no detectadas, se ha decidido utilizar $T_{v_2} = 0,15$.

Por último, en la tercera gráfica de la figura 3.14 se ha representado, para distintos valores del umbral T_{v_3}, el porcentaje de transiciones candidatas que, atendiendo a los criterios de clasificación expuestos en la sección 3.6.3, han sido correctamente etiquetadas como correctas o como falsas según el caso. Atendiendo a los datos obtenidos, se ha decidido utilizar el umbral $T_{v_3} = 5^{\circ}$ aunque, tal y como se puede observar en la mencionada gráfica, para valores de T_{v_3} entre $2,5^{\circ}$ y $7,5^{\circ}$ se obtienen resultados similares (se clasifica correctamente más del 80 % de las transiciones).

El resumen de los resultados obtenidos tras la aplicación de la etapa de clasificación, basada en el análisis del movimiento entre las imágenes de las transiciones candidatas, se ha presentado en la tabla 3.3 en el caso de las transiciones abruptas, y en la tabla 3.4 en el caso de las graduales.

Finalmente, en la tabla 3.5 se presentan los resultados globales del sistema, correspondientes a la detección conjunta de transiciones abruptas y graduales. Observando los resultados de esta tabla se puede apreciar que los porcentajes de *Recall* y *Precision* obtenidos en las cuatro categorías de vídeos utilizadas son muy elevados. El motivo de la presencia de transiciones no detectadas se debe principalmente a situaciones en las que, por culpa de un campo de vectores erróneamente interpretado en la etapa de análisis del movimiento, se ha determinado que existía movimiento entre imágenes pertenecientes a tomas distintas. Por otro lado, la mayor parte de las falsas detecciones son debidas a situaciones en las que los vectores de movimiento han sido insuficientes o de baja calidad como para determinar

coherencia en el desplazamiento entre las imágenes analizadas.

3.7.4. Velocidad del sistema

En último lugar, en esta sección se presentan los resultados relativos a la eficiencia computacional del sistema desarrollado. Dichos resultados se han obtenido haciendo uso de un Intel Core i5 de 2,66 GHz y una memoria RAM de 4 GB.

$H \times W$	Parcial (*fps*)	Total (*fps*)
182×320	61,59	59,76
270×480	56,31	55
360×480	45,82	45,59
360×640	32,08	30,43
576×720	21,09	20,22

Tabla 3.6: Velocidades, en imágenes por segundo (*fps*), obtenidas en secuencias con distinta resolución espacial.

En la tabla 3.6 se muestran, en términos de imágenes analizadas por segundo (*fps*), los resultados correspondientes a varias secuencias con distinta resolución espacial. La segunda columna de esta tabla muestra las velocidades obtenidas con una versión del sistema que no incluye la etapa de análisis del movimiento. Se puede apreciar que, utilizando únicamente las etapas basadas en el análisis de diferencias a nivel de píxel y en el análisis de gradientes, la velocidad alcanzada es muy alta. Prestando atención a los resultados mostrados en la tercera columna de la tabla, los cuales incluyen la etapa de análisis del movimiento, se observa que, gracias a que el análisis del movimiento se realiza únicamente sobre algunos pares de imágenes, el sistema sigue siendo suficientemente rápido, lo cuál hace posible su utilización en aplicaciones que requieran trabajar en tiempo real.

3.8. Conclusiones

En este capítulo se a presentado un novedoso sistema para la segmentación temporal de secuencias de vídeo, el cual es capaz de detectar eficientemente las transiciones entre entre tomas. Mediante un análisis de las diferencias entre imágenes consecutivas a nivel de píxel se detectan las transiciones abruptas existentes y, en paralelo, mediante un análisis de las variaciones en la cantidad de puntos de borde significativos de las imágenes, se identifican las transiciones graduales. Estos dos análisis se han reforzado con una segunda etapa, basada en el análisis del movimiento entre pares de imágenes previamente seleccionadas, que simplifica el problema de la selección de umbrales a la vez que preserva los requisitos computacionales del sistema.

El sistema propuesto ha sido evaluado sobre una amplia gama de secuencias de vídeo, obteniéndose resultados de gran calidad en situaciones poco favorables como, por ejemplo, fondos similares en tomas consecutivas, cambios significativos en el contenido de las secuencias, movimiento de la cámara, grandes objetos móviles en escena y cambios de iluminación.

Además, debido a que la etapa basada en el análisis del movimiento se aplica únicamente sobre algunos pares de imágenes, el sistema ha mostrado ser capaz de funcionar a gran velocidad, lo cual lo hace válido para su utilización en aplicaciones que requieren trabajar en tiempo real.

Capítulo 4

Detección de objetos móviles utilizando mezclas de gaussianas

No guardes nunca en la cabeza
aquello que te quepa en un bolsillo.
Albert Einstein (1879-1955),
científico alemán.

Resumen: Actualmente, debido a los continuos avances tecnológicos, el número de aplicaciones de visión artificial crece muy rápidamente. Estas aplicaciones, como etapa clave para la realización de tareas de alto nivel, precisan de la utilización de estrategias de detección de objetos móviles. Estas estrategias, además de ser capaces de obtener resultados de calidad en cualquier tipo de escenario, han de ser computacionalmente eficientes y deben tener un consumo bajo de memoria. Por este motivo, en esta tesis, se ha desarrollado la estrategia para la detección de objetos móviles que es descrita a lo largo del presente capítulo. Dicha estrategia, basada en el popular método de mezcla de gaussianas, es capaz de adaptar dinámicamente el número de gaussianas utilizadas por cada píxel en cada instante temporal, lo cual hace posible obtener resultados de gran calidad, a la vez que se reduce muy notablemente el coste computacional y los requisitos de memoria asociados al método. De este modo, la estrategia propuesta se muestra idónea para su utilización en un gran número de aplicaciones que, a día de hoy, requieren trabajar en tiempo real. Además, el método propuesto reduce la dependencia de los resultados con los parámetros utilizados por el método original, lo cual facilita su utilización en cualquier tipo de secuencia, independientemente de sus características.

4.1. Introducción

Los avances tecnológicos conseguidos a lo largo de los últimos años han hecho posible la aparición de un gran número aplicaciones de visión artificial que son utilizadas para distintos fines (Elhabian et al., 2008). En estas aplicaciones, la detección de objetos móviles es una

etapa clave y necesaria para la realización de otras tareas de más alto nivel como, por ejemplo, el seguimiento o la clasificación de objetos móviles, el análisis de eventos, etc.

Típicamente, para conseguir resultados precisos y con un número bajo de falsas detecciones se utilizan técnicas de detección basadas en la substracción de fondos. El objetivo de estas técnicas es extraer los fondos presentes en las secuencias analizadas para, de este modo, poder distinguir los objetos que no forman parte de ellos (los objetos móviles). Actualmente, en la literatura, se puede encontrar un amplio abanico de de estrategias de detección basadas en técnicas de substracción de fondos (Benezeth et al., 2009) (Cristani et al., 2010). La calidad de estas estrategias la determinan tanto los resultados que proporcionan como su velocidad de procesamiento (coste computacional) y sus requisitos de memoria (Piccardi, 2005).

Algunas de estas estrategias se centran en maximizar su velocidad y en reducir la memoria que necesitan. Sin embargo, sólo son capaces de proporcionar buenos resultados en secuencias de corta duración en las que no tienen lugar variaciones significativas (Elhabian et al., 2008), resultando poco útiles en presencia de fondos dinámicos. Para resolver estas limitaciones, en los últimos años se han propuesto numerosas estrategias multimodales. Estas estrategias son capaces de modelar varios estados para cada píxel, clasificándolos de forma correcta como parte del fondo de la secuencia incluso cuando pertenecen a regiones no estáticas del fondo.

Dentro del conjunto de estrategias multimodales, el método de mezcla de gaussianas (*Mixture of Gaussians*, *MoG*), propuesto por primera vez en (Grimson y Stauffer, 1999), ha sido uno de los más utilizados a lo largo de los últimos años (Bouwmans et al., 2008). Este método, haciendo uso de múltiples distribuciones gaussianas, trata de modelar las variaciones sufridas por cada píxel para, de ese modo, determinar qué píxeles pertenecen al fondo, y cuáles pertenecen al contenido móvil de las secuencias. De esta forma, si se utiliza un número de gaussianas por píxel suficientemente alto, es capaz de obtener resultados satisfactorios en una gran variedad de secuencias con fondos dinámicos.

Sin embargo, el método de mezcla de gaussianas también tiene algunos inconvenientes que han de ser tenidos en consideración (Piccardi, 2005). El principal de estos inconvenientes es su elevado coste computacional. A la vez que la calidad de los resultados mejora a medida que se utiliza un mayor número de gaussianas para modelar cada píxel, el tiempo de computación del método también se incrementa. Esto supone un serio problema en aplicaciones que requieren trabajar en tiempo real (vídeo-vigilancia, monitorización, etc.) ya que, para ser computacionalmente eficientes, no pueden utilizar un número de gaussianas lo suficientemente alto como para modelar correctamente los fondos de escenarios complejos. Además, otro inconveniente de este método es su alta dependencia con algunos de los parámetros de los que hace uso, lo cual hace que estos parámetros deban tener distinto valor en función de las características de la secuencia bajo análisis.

En este capítulo se presenta una estrategia de substracción de fondos, basada en el método de mezcla de gaussianas, capaz de estimar dinámicamente, en cada instante temporal, el número de gaussianas necesarias para modelar correctamente las variaciones de cada píxel. Asignando pocas gaussianas a los píxeles de las regiones del fondo que sufren pocas variaciones y un mayor número de las mismas únicamente a las regiones más dinámicas se

consigue reducir muy notablemente el coste computacional del método sin que la calidad de los resultados disminuya. Además, el método propuesto reduce la influencia de algunos de los parámetros utilizados por el método original, independientemente de las características de la secuencia analizada, lo cual hace más sencilla su utilización.

El presente capítulo está organizado del siguiente modo: en primer lugar, en la sección 4.2 se describe detalladamente el funcionamiento del método original de mezcla de gaussianas; a continuación, en la sección 4.3 se presenta la alternativa propuesta, basada en la asignación dinámica del número de gaussianas en cada píxel; por último, en las secciones 4.4 y 4.5 se presentan, respectivamente, los resultados obtenidos y las conclusiones del capítulo.

4.2. Substracción de fondos con mezclas de gaussianas

Considérese un píxel cualquiera, p^n, de la imagen I^n, en el instante temporal n de una secuencia de vídeo. Considérese este píxel definido por un vector de D dimensiones, $\mathbf{x}^n \in \mathbb{R}^D$, en el que cada componente es una característica distinta del píxel en cuestión. Haciendo uso de una mezcla de N_K distribuciones gaussianas D-dimensionales es posible modelar su historia reciente y, por lo tanto, determinar su función densidad de probabilidad como (Lee y Lee, 2010):

$$\hat{f}(\mathbf{x}^n) = \sum_{i=1}^{N_K} w_i^n \eta(\mathbf{x}^n; \mu_i^n, \Sigma_i^n) \tag{4.1}$$

donde w_i^n es una estimación del peso de la gaussiana i-ésima de la mezcla en el instante n, μ_i^n es un vector de D componentes que determina la media de dicha gaussiana, Σ_i^n es una matriz de escala de $D \times D$ componentes que determina el ancho de la gaussiana, y η es una función densidad de probabilidad definida como:

$$\eta(\mathbf{x}^n; \mu_i^n, \Sigma_i^n) = \frac{1}{(2\pi)^{\frac{D}{2}} \, |\Sigma_i^n|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} \left(\mathbf{x^n} - \mu_i^n\right)^T \left(\Sigma_i^n\right)^{-1} \left(\mathbf{x^n} - \mu_i^n\right)\right) \tag{4.2}$$

La calidad de los resultados obtenidos con este método depende enormemente del número de gaussianas utilizadas para modelar las variaciones de cada píxel (Wang y Suter, 2005). Si no se utilizan las suficientes, las regiones del fondo que sufran muchas variaciones (aquellas con un elevado número de modos en cada píxel) no podrán ser modeladas correctamente. Por otro lado, si se utilizan demasiadas gaussianas, el resultado es un gran aumento en el coste computacional del sistema y en la cantidad de memoria requerida por el mismo, lo cual dificulta la utilización de este método en aplicaciones que requieren trabajar en tiempo real. Teniendo en cuenta estas consideraciones, normalmente, los trabajos que pueden encontrarse en la literatura utilizan entre 3 y 5 gaussianas por píxel (Elhabian et al., 2008).

Además, para conseguir una reducción importante del coste computacional, sin que la calidad de los resultados se vea gravemente afectada (Benezeth et al., 2009), normalmente se asume la ausencia de correlación en el conjunto de componentes utilizadas, lo que hace

posible utilizar matrices de escala diagonales:

$$\Sigma_i^n = \text{diag}((\sigma_{i,1}^n)^2, (\sigma_{i,2}^n)^2 \dots (\sigma_{i,D}^n)^2) \tag{4.3}$$

en las que cada elemento es la varianza correspondiente a cada una de las D componentes utilizadas. Aplicando estas matrices de escala diagonales, la función densidad de probabilidad definida en la ecuación 4.1 puede escribirse como:

$$\hat{f}(\mathbf{x}^n) = \frac{1}{(2\pi)^{\frac{D}{2}}} \sum_{i=1}^{N_K} w_i^n \prod_{j=1}^{D} \frac{1}{(\Sigma_i^n(j,j))^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^n(j) - \mu_i^n(j))^2}{\Sigma_i^n(j,j)}\right) \tag{4.4}$$

Para determinar si un píxel pertenece al fondo o al primer plano de una secuencia se aplican tres etapas de análisis (Stauffer y Grimson, 2002a). En primer lugar, analizando la distancia entre el píxel y sus gaussianas asociadas, se trata de identificar la gaussiana que modela el valor actual del píxel. En la segunda etapa se actualizan los parámetros de las gaussianas. Por último, dependiendo del número de gaussianas necesarias para modelar el píxel y de la distancia entre estas gaussianas y el píxel, se determina si éste pertenece al fondo o a alguno de los objetos móviles de la secuencia analizada. En las siguientes secciones se describe detalladamente cada una de estas tres etapas.

4.2.1. Identificación de las gaussianas

El primer paso en el análisis de una nueva imagen consiste en analizar la capacidad de las gaussianas para representar los valores de los píxeles de esa imagen. Se considera que un píxel, p^n, está siendo adecuadamente representado por las gaussianas que tiene asociadas si, para alguna de las gaussianas, la distancia entre el píxel y su media es inferior a $2,5$ veces su desviación típica (Bouwmans et al., 2008):

$$\left[\sum_{j=1}^{D} \frac{(x^n(j) - \mu_i^n(j))^2}{\Sigma_i^n(j,j)}\right]^{\frac{1}{2}} < 2,5 \ / \ 1 \leq i \leq N_K \tag{4.5}$$

Si se da el caso en el que ninguna de las gaussianas asociadas a un píxel satisface esta condición, se determina que ninguna de ellas está lo suficientemente próxima al valor actual del píxel como para representarlo adecuadamente. En estos casos se elimina la gaussiana con el menor factor de peso asociado, creándose una nueva en su lugar. La nueva gaussiana se centrará en el valor actual del píxel, $\mu_i = \mathbf{x}^n$, su peso se inicializará como $w_i = w_0$ y su matriz de escala como $\mathbf{\Sigma}_i = \mathbf{\Sigma}_0$, siendo w_0 y $\mathbf{\Sigma}_0$ valores establecidos por el usuario, normalmente con un valor bajo en el caso del peso y un valor alto en de la matriz de escala (Zang y Klette, 2006).

4.2.2. Actualización de las gaussianas

Localizadas las gaussianas que verifican la condición que acaba de ser descrita, el siguiente paso consiste en actualizar los parámetros que las definen: su peso, su media y su matriz

de escala. Esta actualización se lleva a cabo mediante las siguientes expresiones:

$$w^n = (1 - \alpha) \cdot w^{n-1} + \alpha \tag{4.6}$$

$$\mu^n = (1 - \rho^n) \cdot \mu^{n-1} + \rho^n \cdot \mathbf{x}^n \tag{4.7}$$

$$\mathbf{\Sigma}^n = (1 - \rho^n) \cdot \mathbf{\Sigma}^{n-1} + \rho^n \cdot (\mathbf{x}^n - \mu^n) \cdot (\mathbf{x}^n - \mu^n)^T \tag{4.8}$$

en las que α es un parámetro fijo, comprendido entre 0 y 1, que determina la velocidad de actualización de las gaussianas, y ρ es un factor que determina la evolución de cada distribución gaussiana y que se define como:

$$\rho^n = \alpha \cdot \prod_{j=1}^{D} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^n(j) - \mu^n(j))^2}{\mathbf{\Sigma}^n(j,j)}\right) \tag{4.9}$$

Si el valor actual del píxel está lejos de la media de la gaussiana a la que ha sido asociado en la etapa de identificación, el valor de ρ será bajo, lo que reducirá la influencia del valor actual del píxel sobre los parámetros de la gaussiana actualizada. Por el contrario, si el valor del píxel está próximo a la media de esta gaussiana, el píxel tendrá una mayor influencia en la actualización de sus parámetros.

Una vez han sido actualizados los parámetros de las gaussianas que han satisfecho la expresión 4.5, se lleva a cabo la actualización de los factores de peso del resto de las gaussianas. A diferencia de la actualización correspondiente a las gaussianas asociadas a los valores de los píxeles de la imagen actual, cuyo peso se actualiza mediante la ecuación 4.6, el peso del resto de gaussianas se actualiza como:

$$w^n = (1 - \alpha) \cdot w^{n-1} \tag{4.10}$$

Finalmente, en un último paso, los pesos de todas las gaussianas asociadas a cada píxel se normalizan de forma que su suma sea igual a 1:

$$w_i^n = \frac{w_i^n}{\sum_{i=1}^{N_K} w_i^n} \;/\; 1 \leq i \leq N_K \tag{4.11}$$

En el método descrito, la elección de un valor adecuado para el parámetro α tiene una gran relevancia, ya que determina la velocidad de actualización de las gaussianas (White y Shah, 2007), la cual tiene una gran influencia en la calidad de los resultados obtenidos. Si su valor es demasiado bajo las gaussianas se actualizarán lentamente, y esto hará que aumente el número de falsas detecciones en situaciones en las que el fondo de la secuencia sufra cambios rápidos. Por otro lado, si el valor de α es demasiado elevado, la actualización de las gaussianas será muy rápida, lo cual resultará en la errónea clasificación, como píxeles pertenecientes a objetos estáticos, de los píxeles pertenecientes a objetos móviles que se desplacen lentamente o que permanezcan estáticos durante periodos cortos de tiempo. Por lo tanto, dependiendo del tipo de secuencia analizada y de los requisitos de la aplicación

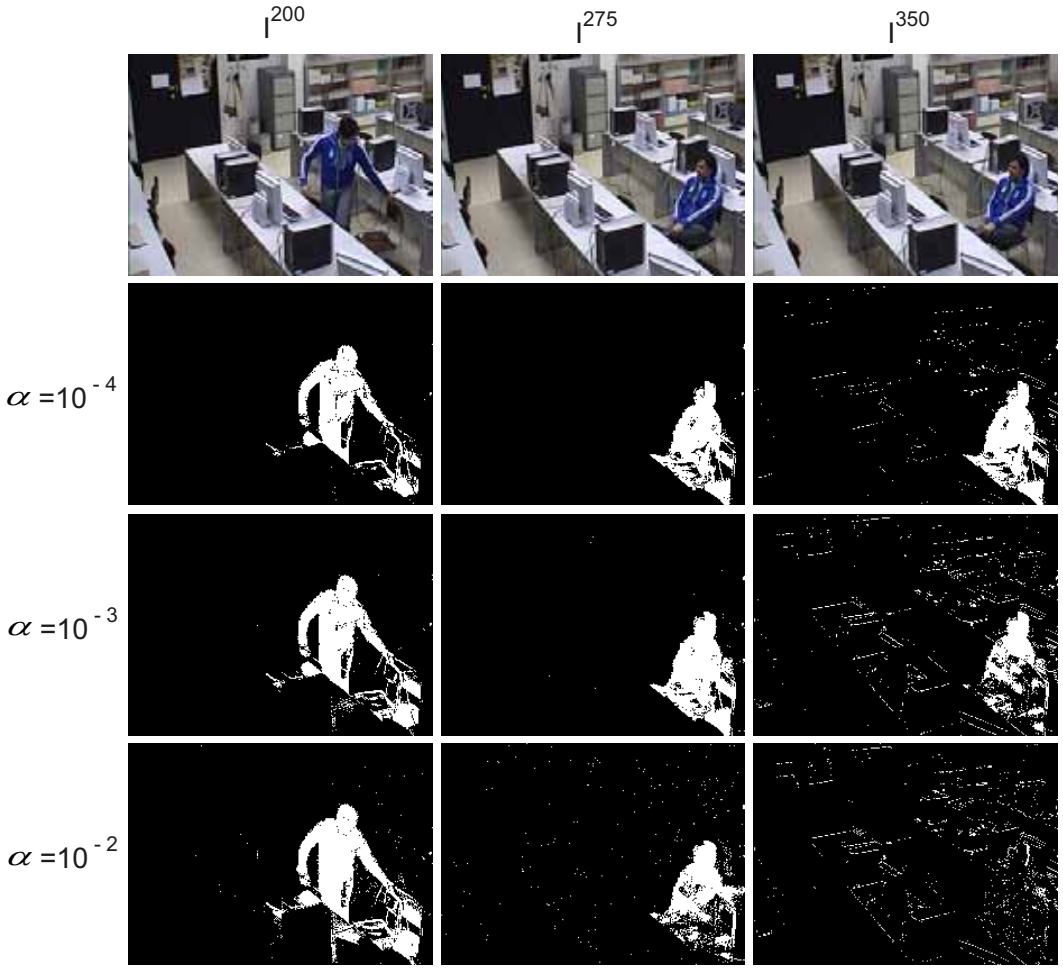


Figura 4.1: Influencia del parámetro α a lo largo de una secuencia en la que un objeto móvil se queda parado.

en la que se esté utilizando el método de detección descrito (reducir el número de falsos positivos o la cantidad de objetos móviles no detectados), el valor asignado a este parámetro debe ser mayor o menor.

En las figuras 4.1 y 4.2 se muestran algunos resultados obtenidos, haciendo uso de distintos valores de α, a lo largo de dos secuencias en las que la velocidad de actualización de las gaussianas influye enormemente en la calidad de los resultados. En la figura 4.1 se presentan los resultados en el caso de una secuencia en la que un objeto móvil se queda parado durante un periodo considerable de tiempo. Se puede observar que, a medida que se utiliza un valor de α más alto, el periodo de tiempo en el que el objeto móvil está siendo correctamente detectado se reduce notablemente. Por otro lado, en el ejemplo de la figura 4.2, se muestran los resultados correspondientes a una secuencia en la que el fondo sufre un cambio rápido y permanente, debido a la substracción de un elemento del mismo por parte

Figura 4.2: Influencia del parámetro α a lo largo de una secuencia en la que el fondo sufre un cambio rápido y permanente.

de un individuo. Los resultados obtenidos para esta secuencia permiten observar que para menores valores de α el tiempo necesario para actualizar el cambio sufrido en el fondo es mayor.

Otro de los aspectos que se puede destacar de los resultados obtenidos en el análisis de estas dos secuencias es el aumento de ruido en las detecciones a medida que se utiliza un α mayor. Esto se debe a que cuanto mayor es α, más rápidamente se estrechan las gaussianas y, consecuentemente, al cubrir un menor rango de posibles valores de los píxeles, modelan menos variaciones ruidosas.

4.2.3. Clasificación de los píxeles

Tras las actualización las gaussianas, en la última etapa del método descrito, los píxeles son clasificados como estáticos (parte del fondo de la secuencia) o como dinámicos (parte de objetos los móviles). Para llevar a cabo esta clasificación, para cada píxel, se siguen los

pasos que se describen a continuación.

1. Las gaussianas asociadas al píxel se ordenan de mayor a menor, en función del ratio:

$$r_G = \frac{w_i^n}{\prod_{j=1}^{D}(\Sigma_i^n(j,j))^{\frac{1}{2}}} \tag{4.12}$$

   Por definición, un píxel perteneciente al fondo de una secuencia debe poseer valores muy similares durante periodos prolongados de tiempo (Stauffer y Grimson, 2002b) o, lo que es lo mismo, las gaussianas que modelen sus variaciones deberán tener un peso elevado y una desviación típica pequeña. Por lo tanto, ordenando las gaussianas en el orden decreciente determinado por el ratio r_G, estarán siendo ordenadas en función de cómo de bien verifiquen esta definición.

2. Se selecciona el número mínimo de gaussianas, N_G, cuyos factores de peso sumados superen el valor de un umbral, T_G:

$$N_G = \arg\min_j \left(\sum_{i=1}^{j} w_i^n > T_G \right) \tag{4.13}$$

3. Si el píxel bajo análisis pertenece a alguna de estas N_G gaussianas, lo cual se determina comparando el valor del píxel con las medias de las gaussianas mediante la ecuación 4.3, se clasifica como estático. En caso contrario se clasifica como dinámico.

En este proceso de decisión, el umbral T_G determina la porción mínima de los datos que se tiene en cuenta para decidir si un píxel pertenece al fondo de una secuencia (Ha y Lee, 2010). Por lo tanto, influye directamente en la capacidad del método para clasificar correctamente los píxeles del fondo con variaciones multimodales. Si su valor es demasiado bajo, al considerarse pocas gaussianas, se reducirá la capacidad del método para representar las variaciones multimodales de los píxeles del fondo. Por otro lado, si su valor es demasiado elevado, los píxeles pertenecientes a los objetos móviles pueden ser erróneamente clasificados como parte del fondo de la secuencia. La figura 4.3 muestra los resultados obtenidos, utilizando distinto valor de T_G, sobre varias imágenes de una misma secuencia. Observando estos resultados se puede apreciar que, utilizando un T_G de menor valor, el número de píxeles erróneamente etiquetados como móviles es mayor (píxeles del árbol que aparece en primer plano). Sin embargo, para mayores valores de T_G ha aumentado en número de píxeles móviles no detectados.

4.3. Detección de objetos móviles con mezclas variables de gaussianas

El método de mezcla de gaussianas previamente descrito proporciona resultados de gran calidad en una amplia variedad de escenarios: secuencias con fondos dinámicos, cambios de iluminación en la escena, o secuencias con una cantidad significativa de ruido. Sin embargo, la calidad de estos resultados está muy condicionada por la adecuada elección de los valores de algunos parámetros:

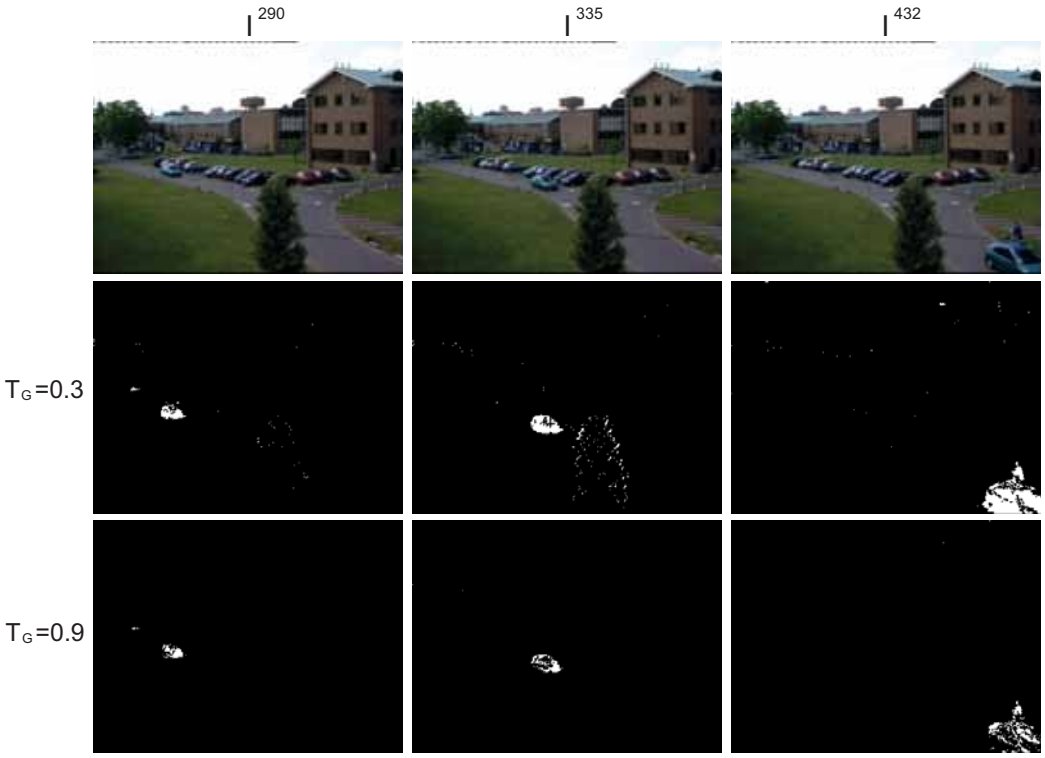


Figura 4.3: Influencia del umbral T_G en la calidad de los resultados.

- α: Determina la velocidad de actualización de las gaussianas. Por un lado, si su valor es demasiado alto, los objetos móviles que se desplazan lentamente no serán adecuadamente detectados. Por otro lado, si su valor es demasiado bajo, en secuencias en las que el fondo sufra variaciones rápidas aumentará el número de falsas detecciones.
- T_G: Determina el número de gaussianas utilizadas para decidir si un píxel es estático o dinámico. Si su valor es demasiado bajo se reducirá la capacidad del método para representar las variaciones multimodales del fondo. Sin embargo, si su valor es demasiado alto aumentará el número de falsas detecciones.
- N_K: Establece el número de gaussianas utilizadas para estimar la función densidad de probabilidad de cada píxel. Pocas gaussianas impedirán representar correctamente las variaciones multimodales del fondo, mientras que demasiadas gaussianas supondrán un importante aumento del coste computacional del método, dificultando su utilización en aplicaciones que requieran trabajar en tiempo real.

En esta sección se describe una novedosa y eficiente estrategia de detección de objetos móviles, basada en el método de mezcla de gaussianas descrito en la sección 4.2, capaz de seleccionar dinámicamente el número de gaussianas utilizadas para modelar cada píxel en cada instante de tiempo, en función de sus necesidades. De este modo, asignando mayor o menor número de gaussianas a cada píxel, en función de las variaciones que sufre, se consigue una más que apreciable reducción del número medio de gaussianas utilizadas en

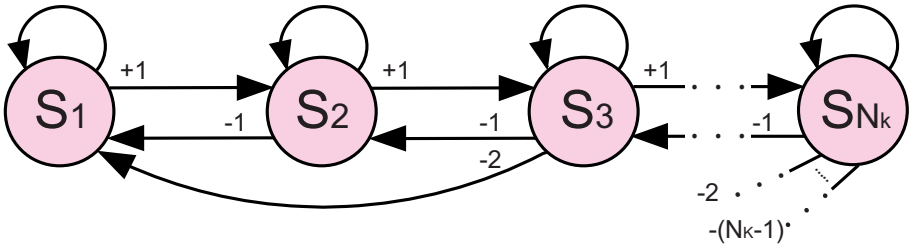


Figura 4.4: Diagrama de transiciones para cada píxel.

cada imagen. Si se tiene en cuenta la cantidad de operaciones que se realizan sobre cada una de ellas en el análisis de cada imagen, esta reducción de las gaussianas se traduce en un importante ahorro computacional, lo cual facilita su integración en aplicaciones que requieren trabajar en tiempo real.

Además, como se verá a lo largo de la presente sección, gracias a la adaptación dinámica del número de gaussianas en cada píxel, el método propuesto permite reducir la influencia de los parámetros α y T_G en los resultados, lo cual hace más sencilla la utilización del método, independientemente del tipo de secuencia sobre la que se aplique.

4.3.1. Descripción de la estrategia

El objetivo de la estrategia desarrollada consiste en determinar, de forma automática, el número mínimo de gaussianas que necesita cada píxel en cada instante temporal, de forma que sus variaciones puedan ser adecuadamente modeladas.

En la Figura 4.4 se muestra el diagrama de transiciones que describe los posibles cambios de estado de un píxel. Este diagrama consta de un conjunto de N_K estados, definidos como $\{S_i\}_{i=1}^{N_K}$. Cada uno de estos estados se caracteriza por el número de gaussianas asociadas a cada píxel en cada instante, siendo los píxeles que se encuentren en el estado S_i, aquellos que tengan i gaussianas asociadas. Los números que aparecen sobre las fechas representan el número de gaussianas que son añadidas (en el caso de los números positivos) o eliminadas (en el caso de los negativos) en las transiciones entre los estados.

En función de las variaciones de cada píxel en su historia reciente le será asignado un mayor o menor número de gaussianas (con un mínimo de 1 gaussiana en el caso de píxeles muy estáticos y un máximo de N_K gaussianas en el caso de píxeles que sufran importantes variaciones). Cuando se determine que un píxel precisa de más gaussianas para modelar adecuadamente sus variaciones, siempre que no haya alcanzado el número máximo de gaussianas posibles, pasará del estado S_i al S_{i+1}. Por contra, si se determina que, para un píxel dado, se están utilizando más gaussianas de las necesarias, dicho píxel pasará del estado S_i al S_{i-q}, siendo q el número de gaussianas que se hayan considerado no necesarias.

4.3.2. Implementación del método

En esta sección se describe detalladamente el modo de aplicación de la estrategia propuesta, así como las condiciones que han de darse en un píxel para que sufra un cambio de estado y los criterios a considerar para determinar si ese píxel es estático o dinámico.

Sea p^n un píxel cualquiera de la imagen I^n, correspondiente al instante temporal n. En ese preciso instante, este píxel tendrá asociado un conjunto de k^n gaussianas, $\{G_i^n\}_{i=1}^{k^n}$, y un contador asociado a cada gaussiana, C_i^n. Esta asociación se define como:

$$p^n \leftrightarrow \{G_i^n(\mu_i^n, \Sigma_i^n, \omega_i^n), C_i^n\}_{i=1}^{k^n} \tag{4.14}$$

donde $k^n \in [1, N_K]$ determina el número de gaussianas asociadas al píxel, μ_i^n es la media de la gaussiana i-ésima del conjunto, Σ_i^n es su matriz de escala y ω_i^n es su peso asociado.

Siempre que se crea una nueva gaussiana su contador asociado se inicializa con un valor inicial prefijado, C_0, el cual determina el tiempo máximo que ha de transcurrir para que dicha gaussiana, en el caso de no ser necesaria, sea eliminada. Los valores de los contadores serán actualizados en función de los resultados obtenidos en la etapa de identificación de las gaussianas.

Al comenzar el análisis, en la primera imagen de la secuencia, a cada píxel se le asigna una única gaussiana, $k^1 = 1$. Esta gaussiana se inicializa con un valor medio igual al valor del píxel, $\mu_1^1 = x^1$, con un peso w_0 y con una matriz de escala Σ_0. Además, los contadores de todas estas gaussianas se inicializan con el valor C_0. En esta primera imagen, por lo tanto, todos los píxeles estarán en el estado S_1 y serán etiquetados como píxeles del fondo de la secuencia (píxeles estáticos).

En imágenes sucesivas, sobre cada píxel, se aplicarán las etapas de identificación, actualización y clasificación descritas en las secciones 4.2.1, 4.2.2 y 4.2.3, respectivamente, teniendo en cuenta las siguientes consideraciones en la etapa de identificación de las gaussianas. Dependiendo del resultado de la ecuación 4.5 aplicada sobre las gaussianas de cada píxel, se llevará a cabo una de las siguientes acciones:

- Si alguna de las gaussianas ha satisfecho la ecuación 4.5:
  - Se actualiza el contador de dicha gaussiana con el valor C_0.
  - Se reduce en una unidad el contador del resto de gaussianas asociadas al píxel: $C_i = C_i - 1$.
  - Las q gaussianas cuyos contadores hayan llegado a 0 son eliminadas, pasando el píxel del estado S_{k^n} al S_{k^n-q}.
- Si ninguna gaussiana ha satisfecho la ecuación 4.5:
  - Se reduce en una unidad el contador de todas las gaussianas asociadas al píxel: $C_i = C_i - 1$.
  - Las q gaussianas cuyos contadores hayan llegado a 0 son eliminadas, pasando el píxel del estado S_{k^n} al S_{k^n-q}.
  - Si el píxel no ha alcanzado su máximo posible de gaussianas ($k^n < N_K$):
    - Se añade una nueva gaussiana centrada en el valor actual del píxel, pasando éste al estado S_{k^n+1}.

  - El contador de dicha gaussiana se inicializa con el valor C_0.
- Si el píxel está en el estado S_{N_K}:
  - La gaussiana cuyo contador tenga el menor valor es sustituida por una nueva del modo descrito en la sección 4.2.1. Consecuentemente, el píxel permanece en el estado S_{N_K}.
  - El contador de la nueva gaussiana se inicializa con el valor C_0.

Tras la identificación de las gaussianas, se pasa a la etapa de actualización sus los parámetros. Este proceso de actualización se lleva a cabo exactamente del modo expuesto en la sección 4.2.2. Sin embargo, es importante mencionar que, gracias a la utilización del contador asociado a cada gaussiana, el cual determina el tiempo máximo de supresión de una gaussiana que no está siendo actualizada, el método propuesto permite reducir la dependencia del parámetro α con la velocidad de actualización del fondo. Mientras que en el método original el valor de α determina la velocidad de actualización del fondo de las secuencias, en el método propuesto, si se utiliza un valor de α suficientemente bajo, esta velocidad de actualización va a depender del valor asignado al parámetro C_0. Los valores adecuados para estos dos parámetros se justifican en la sección 4.3.3.

En último lugar, al igual que en el método original, para cada píxel se seleccionan las N_G gaussianas con mayor ratio r_G y se decide si el píxel es parte de un objeto estático o de uno dinámico.

4.3.3. Selección de parámetros

Los parámetros que utiliza la estrategia propuesta y que deben ser manualmente seleccionados por el usuario son los que se resumen a continuación:

- Σ_0: Matriz de escala asignada a las gaussianas en el momento de su creación.
- w_0: Peso asignado a las gaussianas en el momento de su creación.
- α: Parámetro que determina la velocidad de actualización de las gaussianas.
- C_0: Valor inicial del contador asociado a cada gaussiana.
- T_G: Umbral que determina el número de gaussianas utilizadas para clasificar a un píxel como estático o como dinámico.

En las siguientes secciones se discute sobre los valores más adecuados para estos parámetros y sobre su influencia en los resultados proporcionados por el método propuesto.

4.3.3.1. Inicialización de las gaussianas

Tal y como se ha mencionado en el apartado 4.2.1, cuando se crea una nueva gaussiana, su matriz de escala inicial, Σ_0, debe tener valores elevados, mientras que su peso inicial, w_0, debe tener un valor bajo (Zang y Klette, 2006). De esta forma, en comparación con el resto de gaussianas asociadas al píxel en cuestión, el valor de r_G de la nueva gaussiana será el menor y, por lo tanto, hasta que no haya sido actualizada varias veces, no será utilizada en la etapa de clasificación del píxel.

Sin embargo, dependiendo del valor asignado a α, la utilización de unos u otros valores en estos parámetros influye en la calidad de los resultados ya que, cuanto menor es el valor

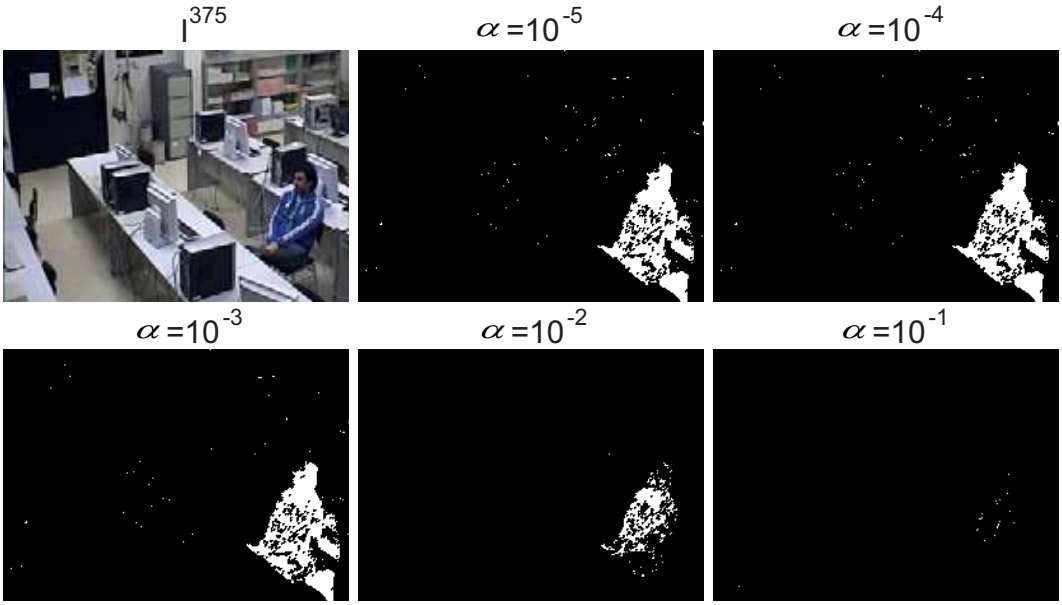


Figura 4.5: Influencia de la elección de α en relación con el valor de C_0.

de α, mayor es el número de actualizaciones que requiere una nueva gaussiana para alcanzar un valor de r_G que permita su utilización en el proceso de clasificación del píxel.

Teniendo en cuenta estas consideraciones, a partir del valor de α utilizado, cuya elección se justifica más adelante, se ha decidido utilizar un peso inicial $w_0 = 0{,}01$ y una matriz de escala cuyos D elementos sean iguales al 5 % del valor máximo de cada una de las D componentes utilizadas. Para valores en el entorno de los elegidos, dado que las condiciones exigidas para estos parámetros siguen verificándose, se ha comprobado que las detecciones obtenidas no sufren cambios apreciables.

4.3.3.2. Velocidad de actualización

En el método original de mezcla de gaussianas, el valor de α determina lo rápido que se actualiza el modelo del fondo, condicionando la cantidad de actualizaciones que necesita una gaussiana recién generada para ser considerada más relevante que el resto de gaussianas existentes o, lo que es lo mismo, pasar a tener un ratio r_G mayor que el de las demás gaussianas.

Sin embargo, en el método propuesto, las gaussianas que dejen de ser actualizadas durante un periodo de tiempo prolongado son eliminadas y, por lo tanto, el tiempo máximo durante el que una gaussiana que no está siendo actualizada puede influir en el resultado de la etapa de clasificación también depende del valor de C_0.

Es necesario tener en cuenta que, aunque con C_0 se puede controlar el tiempo máximo de actualización de las regiones del fondo, el tiempo mínimo de actualización de estas regiones depende, en principio, del valor de α. En la figura 4.5 se muestra un ejemplo con

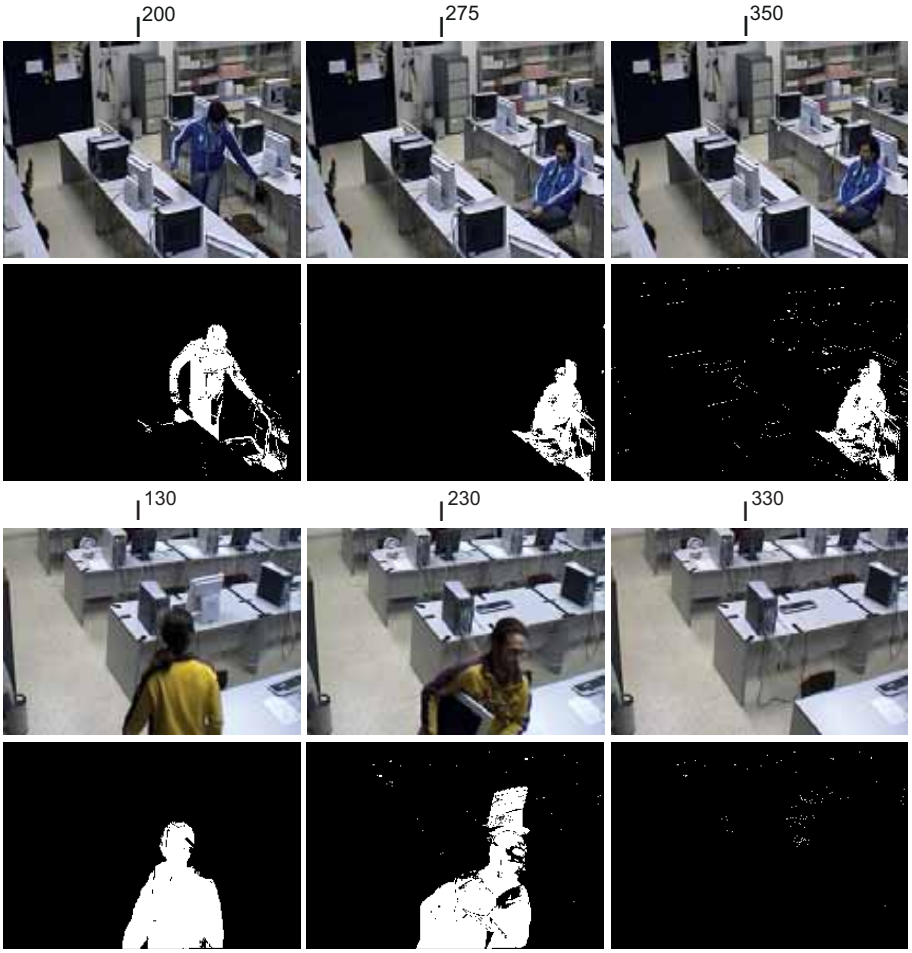


Figura 4.6: Influencia de α en secuencias con distintos requisitos de actualización del fondo.

los resultados obtenidos, utilizando distintos valores de α y un contador inicial con valor $C_0 = 150$, en el caso de una imagen en la que aparece un objeto móvil que se ha quedado parado aproximadamente en la imagen I^{230}. Observando los resultados de esta figura se aprecia que, con $\alpha = 10^{-2}$ y $\alpha = 10^{-1}$, la velocidad de actualización de las gaussianas ha hecho que el fondo se actualice antes de que el contador de las gaussianas del objeto móvil parado llegase a 0. Sin embargo, utilizando valores inferiores de α, en el instante analizado, la actualización de los parámetros de las gaussianas todavía no ha dado lugar a la actualización del fondo, por lo que será en el momento en el que los contadores de las gaussianas lleguen a 0 cuando el fondo sea actualizado. Por lo tanto, si se utiliza un valor de α suficientemente bajo, el valor asignado a C_0 determinará el tiempo aproximado de actualización del fondo.

Para adaptar la velocidad de actualización del fondo a los criterios de actualización descritos en la sección A.2.1 del apéndice A, donde se detalla el modo de generación de las detecciones de referencia utilizadas para evaluar la calidad de las estrategias de detección

descritas a lo largo de esta tesis, se ha decidido asignar al contador un valor inicial de $C_0 = 150$. De esta forma, en situaciones en las que el fondo sufre cambios permanentes, el tiempo de actualización del mismo será de unos 6 segundos (para secuencias con 25 imágenes por segundo). En los experimentos realizados se ha comprobado que utilizando valores de α iguales o inferiores a 10^{-4}, en situaciones en las que el fondo sufre variaciones bruscas, su tiempo aproximado de actualización depende principalmente del valor asignado a C_0. Por otro lado, si α es excesivamente pequeño, las gaussianas no se adaptarán suficientemente rápido a los cambios lentos sufridos por el fondo (Zang y Klette, 2006), incrementándose la cantidad de falsas detecciones en algunas situaciones. Por ese motivo, en los experimentos realizados se ha utilizado $\alpha = 10^{-4}$.

La figura 4.6 muestra los resultados obtenidos tras la aplicación de la estrategia propuesta, con los valores de C_0 y α previamente descritos, sobre dos secuencias en las que la velocidad de actualización del fondo es de gran importancia y en las que, tal y como se puso de manifiesto en los ejemplos mostrados a lo largo de la sección 4.2.2, la calidad de las detecciones obtenidas con el método original de mezcla de gaussianas es muy dependiente del valor de α. Por un lado, la secuencia de la parte superior de la figura presenta un caso en el que un objeto móvil se queda parado permanentemente. Por otro lado, la secuencia de la parte inferior de la figura muestra una situación en la que un objeto del fondo es sustraido por un individuo. Observando los resultados obtenidos se puede apreciar que, a pesar de utilizarse el mismo valor de α, en ambos casos se obtienen resultados satisfactorios: el objeto móvil de la primera secuencia permanece clasificado como tal a pesar de llevar varios segundos parado; y el fondo de la segunda secuencia se ha actualizado poco después de ser modificado.

4.3.3.3. Número de gaussianas

Una vez decididos los valores más adecuados para los parámetros de inicialización y de actualización de las gaussianas, sólo falta decidir qué valor asignar al umbral T_G. El valor de este umbral, recordando lo explicado en la 4.2.3, determina la proporción de los datos representados por las gaussianas que se tienen en cuenta para clasificar a los píxeles como dinámicos o como estáticos (Ha y Lee, 2010). Por lo tanto, el valor asignado a este parámetro influirá en la capacidad del método para representar las variaciones multimodales del fondo.

En el método original de mezcla de gaussianas, si se utiliza un valor demasiado alto se pueden estar teniendo en cuenta los datos asociados a gaussianas que no se correspondan con ninguno de los modos actuales de los píxeles, lo que puede dar lugar un mayor número de píxeles móviles clasificados como parte del fondo (Zang y Klette, 2006).

Sin embargo, en el método propuesto las gaussianas asociadas a cada píxel son exclusivamente aquellas que recientemente han satisfecho la condición de identificación determinada por la ecuación 4.5. De ese modo, se reduce notablemente el número de casos en los que, a pesar de utilizarse la mayor parte de las gaussianas en el proceso de clasificación, los píxeles de los objetos móviles son erróneamente clasificados como parte del fondo. Por lo tanto, dado que asignando a T_G un valor elevado se reduce la aparición de falsas detecciones (píxeles estáticos clasificados como móviles) en las zonas del fondo muy multimodales (Zang y Klette, 2006), se ha decidido utilizar el valor $T_G = 0,9$.

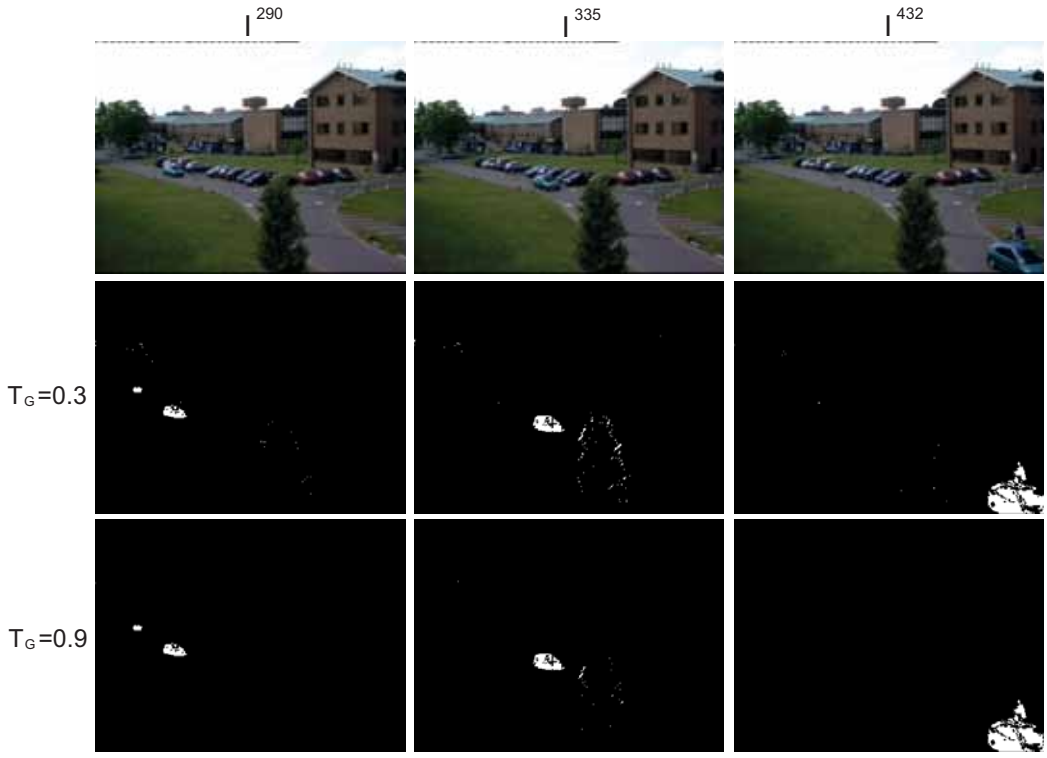


Figura 4.7: Influencia del umbral T_G en la calidad de los resultados.

En la figura 4.7 se muestran los resultados obtenidos con el método propuesto y utilizando distintos valores de T_G. Comparando estos resultados con los mostrados en la figura 4.3 se puede apreciar que, a diferencia que en el método original, utilizando un umbral alto se consigue un resultado con pocas falsas detecciones sin que los objetos móviles sean erróneamente clasificados.

4.4. Resultados

En esta sección se presentan los resultados obtenidos con la estrategia propuesta, tras su aplicación sobre la base de datos descrita en la sección A.2 del apéndice A.

Todas las pruebas se han efectuado sobre un Intel Core i5 de $2,66$ GHz con una memoria RAM de 4 GB. Para su realización se utilizando únicamente la información de color RGB de los píxeles por lo que, en las ecuaciones presentadas en este capítulo, el número de componentes utilizadas ha sido de $D = 3$. Además, se ha establecido un número máximo de $N_K = 5$ gaussianas por píxel. En cuanto al resto de parámetros, se han seleccionado los valores justificados a lo largo de la sección 4.3.3.

A lo largo de esta sección se analiza tanto la eficiencia computacional del método como la calidad de los resultados que proporciona. Además, todos los resultados se han comparado con los resultantes de la aplicación del método clásico de mezcla de gaussianas, en el que

se ha utilizado la configuración de parámetros propuesta en (Atev et al., 2005).

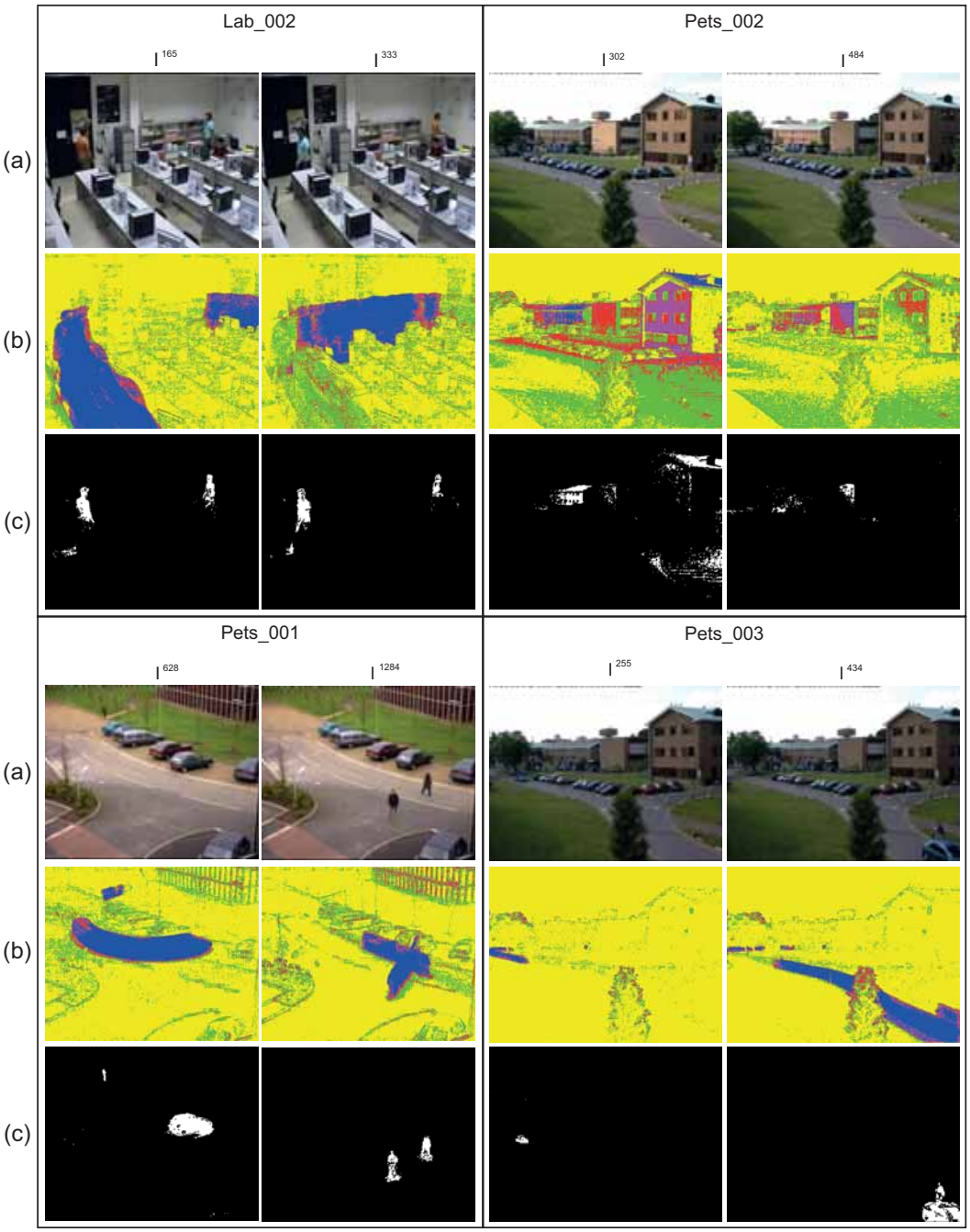


Figura 4.8: (a) Imágenes originales. (b) Máscaras representando el número de gaussianas utilizadas en cada píxel. (c) Detecciones resultantes de la aplicación del método propuesto.

En primer lugar, en la figura 4.8, se presentan algunos resultados obtenidos tras la

	$\mathbf{S_1}(\%)$	$\mathbf{S_2}(\%)$	$\mathbf{S_3}(\%)$	$\mathbf{S_4}(\%)$	$\mathbf{S_5}(\%)$	$\mathbf{R_{G_1}}(\%)$
Lab_001	78,73	13,54	2,48	1,80	3,45	72,46
Lab_002	64,82	16,72	2,94	2,80	12,72	63,62
Lab_003	59,12	24,35	5,12	2,66	8,76	64,48
Lab_004	66,00	13,99	3,67	3,68	12,67	63,40
Lab_005	78,31	13,15	2,08	2,16	4,29	71,81
Lab_006	45,79	27,87	6,16	3,94	16,25	56,60
Pets_001	75,44	15,36	2,36	1,18	5,66	70,75
Pets_002	60,30	28,98	7,33	2,75	0,63	69,12
Pets_003	64,58	33,66	0,68	0,30	0,78	72,19
Pets_004	41,81	10,87	10,81	12,33	24,18	46,76
Pets_005	33,77	6,75	10,95	15,12	33,40	38,48
Wall_001	72,82	4,57	5,11	3,26	14,23	63,70
Wall_002	31,46	11,83	9,81	7,80	39,10	37,75
Total	60,05	17,54	5,26	4,97	12,18	61,66

Tabla 4.1: Porcentaje medio del número de píxeles en cada estado (S_i) a lo largo de cada secuencia y porcentaje de reducción del número total de gaussianas utilizadas con respecto al método original (R_{G_1}).

aplicación del método propuesto sobre cuatro secuencias de distintas características. Cada una de las imágenes originales representadas en esta figura (figura 4.8.a) está acompañada de dos imágenes más: una imagen en color (figura 4.8.b) y una imagen binaria con las detecciones obtenidas (figura 4.8.c). En las imágenes de color, cada color muestra el número de gaussianas que tiene asociadas cada píxel en cada instante o, lo que es lo mismo, el estado en el que se encuentra cada píxel:

- Amarillo: Píxeles en el estado S_1.
- Verde: Píxeles en el estado S_2.
- Rojo: Píxeles en el estado S_3.
- Rosa: Píxeles en el estado S_4.
- Azul: Píxeles en el estado S_5.

Prestando atención a las imágenes de color correspondientes a los ejemplos mostrados en esta figura se puede apreciar que, en todas las secuencias analizadas, la mayor parte de los píxeles tienen asociada una única gaussiana, lo cual pone de manifiesto que, en la mayor parte de las regiones, las variaciones de los píxeles no son multimodales, siendo necesarias más gaussianas sólo en algunas regiones:

- Regiones por las que recientemente ha pasado un objeto móvil: en estas regiones, al producirse importantes cambios respecto a los valores que tenían los píxeles del fondo, se genera un alto número de gaussianas para tratar de modelar estas variaciones. Pasado un tiempo, establecido por el valor de C_0, estas gaussianas comienzan a ser eliminadas.
- Regiones del fondo que no permanecen estáticas: zonas que necesitan múltiples gaussianas para modelar sus variaciones como, por ejemplo, las copas de los árboles en la

	Método original (ms)					**Método propuesto (ms)**	$\mathbf{R_{G_2}}$ (%)
	$\mathbf{N_K = 1}$	$\mathbf{N_K = 2}$	$\mathbf{N_K = 3}$	$\mathbf{N_K = 4}$	$\mathbf{N_K = 5}$		
Lab_001	23,10	29,48	31,39	33,55	35,74	27,50	23,06
Lab_002	21,54	27,68	30,49	32,52	36,15	29,25	19,09
Lab_003	8,96	14,85	16,17	17,62	19,80	15,50	21,71
Lab_004	10,36	15,07	16,45	17,53	19,32	15,83	18,05
Lab_005	21,60	27,74	30,56	31,40	35,37	27,97	20,92
Lab_006	7,46	11,04	13,76	14,90	16,46	13,31	19,11
Pets_001	23,29	30,08	32,98	35,56	39,28	31,08	20,88
Pets_002	22,41	30,38	34,10	36,94	41,06	32,05	21,93
Pets_003	23,63	28,41	31,54	33,36	37,52	28,54	23,94
Pets_004	21,52	28,13	31,81	34,08	38,75	32,53	16,05
Pets_005	21,42	28,11	30,96	33,31	37,68	33,83	10,23
Wall_001	7,13	11,35	14,13	15,26	16,12	13,37	17,09
Wall_002	7,20	11,50	14,32	15,20	15,76	13,89	11,85

Tabla 4.2: Tiempos medios de procesamiento por imagen, expresados en milisegundos. Las primeras columnas muestran los resultados obtenidos con el método original y distintos valores de N_K. La penúltima columna contiene los resultados obtenidos con el método propuesto. La última columna muestra el porcentaje de mejora obtenido con respecto al método original y $N_K = 5$.

secuencia *Pets*_002 o la vegetación en la secuencia *Pets*_001.

- Bordes de los objetos, debido a que sufren más variaciones (por pequeños movimientos de la cámara, por el resultado de la codificación y por su alta sensibilidad a los cambios de iluminación) y, por lo tanto, su varianza es muy superior al de las zonas homogéneas.
- Regiones sometidas a fuertes cambios de iluminación: ejemplo mostrado mediante la secuencia *Pets*_002.

En la tabla 4.1 se presenta un resumen con el porcentaje medio de píxeles en cada uno de sus posibles estados y para todas las secuencias de la base de datos utilizada. La última columna de esta tabla muestra el porcentaje en el que se ha reducido el número de gaussianas con respecto al caso de utilizar el método original con $N_K = 5$ gaussianas en cada píxel durante toda la secuencia. Este porcentaje se ha obtenido aplicando la siguiente expresión:

$$R_{G_1} = \frac{N_{G_1} - N_{G_2}}{N_{G_1}} \tag{4.15}$$

donde N_{G_1} es el número de gaussianas utilizadas en el método original y N_{G_2} es el número de gaussianas utilizadas por el método propuesto. La última fila de la tabla contiene los porcentajes correspondientes al análisis del número de gaussianas utilizadas en el conjunto de todas las secuencias de la base de datos. Estos resultados muestran que aplicando la estrategia propuesta se consigue reducir muy significativamente el número de gaussianas utilizadas (en media, un 61,66 %), lo cual supone una importante reducción del número

de operaciones realizadas en el análisis de las secuencias y, por lo tanto, un gran ahorro computacional.

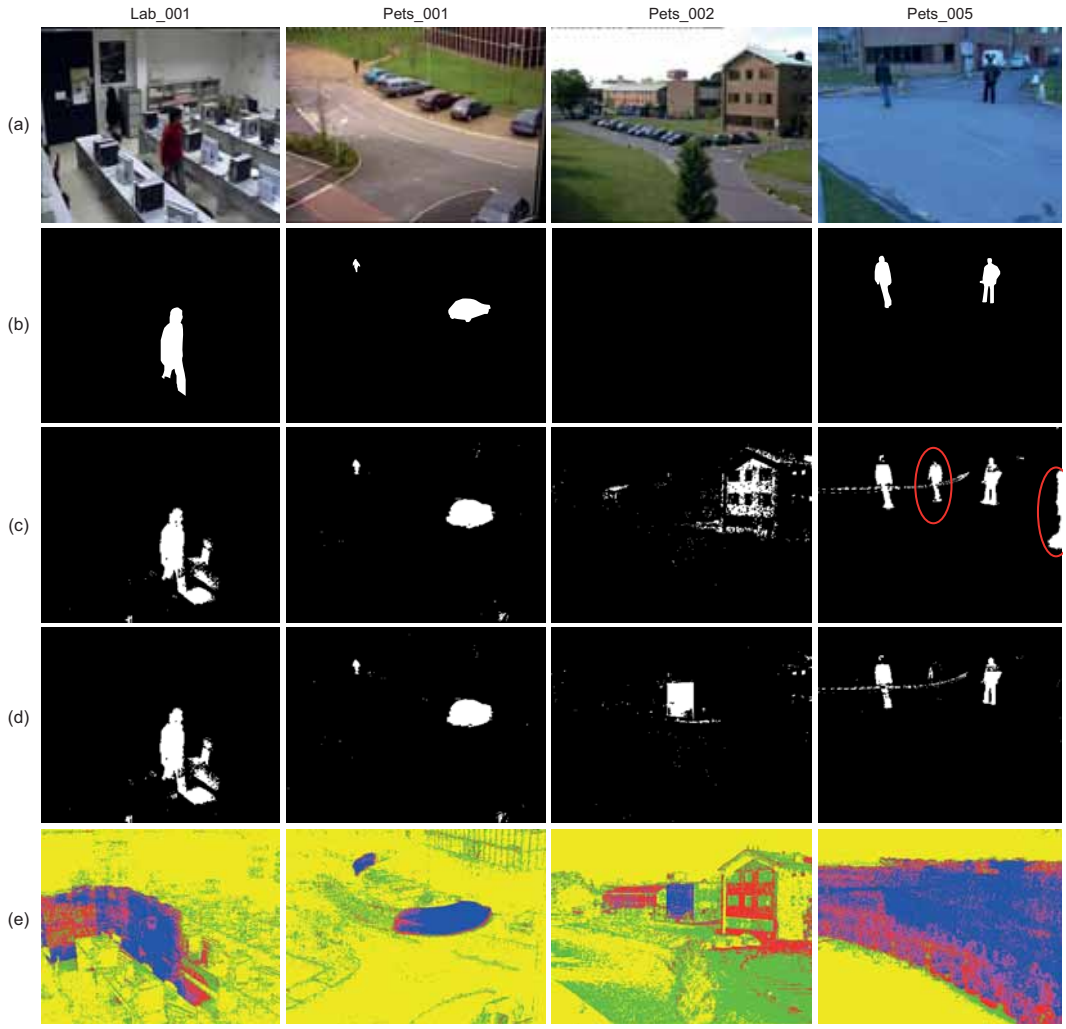


Figura 4.9: Análisis cualitativo de la calidad obtenida con el método propuesto y con el método original de mezcla de gaussianas. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con el método original, utilizando $N_K = 5$ gaussianas por píxel. (d) Detecciones obtenidas con el método propuesto. (e) Máscaras representando el número de gaussianas utilizadas en cada píxel.

Por otro lado, en la tabla 4.2, se muestra información relativa a la velocidad de procesamiento obtenida en el análisis de todas las secuencias utilizadas. En las primeras columnas de la tabla se muestran los tiempos medios de procesamiento por imagen, expresados en milisegundos, obtenidos con el método original de mezcla de gaussianas y para distintos valores de N_K. La penúltima columna muestra los tiempos medios de procesamiento por imagen obtenidos con el método propuesto. La última columna de la tabla presenta el por-

	Método original			**Método propuesto**		
	Recall	Precision	**F**	Recall	Precision	**F**
Lab_001	75,07	82,95	78,81	73,19	83,92	78,19
Lab_002	77,23	83,85	80,40	76,39	84,37	80,18
Lab_003	79,17	46,35	58,47	77,56	65,72	71,15
Lab_004	84,76	57,16	68,28	83,60	69,24	75,75
Lab_005	69,92	83,76	76,22	69,03	84,09	75,82
Lab_006	91,28	52,58	66,72	90,20	58,98	71,63
Pets_001	91,37	51,13	65,57	87,54	71,77	78,87
Pets_002	100	0	0	100	0	0
Pets_003	79,04	67,57	72,86	81,49	69,21	74,85
Pets_004	76,49	64,52	69,99	74,37	81,94	77,97
Pets_005	83,84	81,48	82,64	82,34	90,28	86,13
Wall_001	95,44	43,87	60,11	95,49	43,83	60,09
Wall_002	95,80	8,65	15,87	95,55	6,72	12,56
Total	87,68	18,07	29,97	86,78	20,84	33,60

Tabla 4.3: Resultados cualitativos correspondientes a las detecciones obtenidas mediante el método original de mezcla de gaussianas (con $K = 5$) y mediante el método propuesto.

centaje en el que se ha conseguido reducir el tiempo de computación con respecto al caso de utilizar el método original con $K = 5$ gaussianas. Este porcentaje se ha obtenido aplicando la siguiente expresión:

$$R_{G_2} = \frac{T_{G_1} - T_{G_2}}{T_{G_1}} \tag{4.16}$$

donde T_{G_1} es el tiempo mostrado en la sexta columna de la tabla (obtenido mediante el método original con $K = 5$) y T_{G_2} es el tiempo mostrado en la séptima columna de la tabla (obtenido con nuestro método). Los resultados mostrados en esta tabla muestran que aplicando el método de detección propuesto se ha conseguido una mejora computacional muy significativa: entre un $11,85\,\%$ y un $23,94\,\%$, dependiendo de las características y del contenido móvil de la secuencia analizada.

En último lugar se ha evaluado la calidad de los resultados obtenidos y se ha comparado con la resultante de la aplicación del método original de mezcla de gaussianas. Algunos de estos resultados, obtenidos para cuatro secuencias de distintas características, pueden verse en la figura 4.9. En todos los ejemplos mostrados en esta figura se puede apreciar que, con el método propuesto, la calidad de las detecciones es muy similar a la obtenida con el método original. Sin embargo, los resultados proporcionados por nuestro método contienen una cantidad de falsas detecciones ligeramente superior a la del método original. Estas falsas detecciones se deben a los instantes en los que los píxeles del fondo necesitan incrementar su número de gaussianas para ser modelados adecuadamente, momento en el que estos píxeles son clasificados como dinámicos. Por otro lado, observando los resultados correspondientes a las secuencias $Pets\_002$ y $Pets\_005$, se pueden apreciar algunas mejoras

de nuestra estrategia frente a la original. En la primera de estas dos secuencias, gracias a la utilización del parámetro C_0, el cual limita el tiempo máximo de actualización del fondo, conseguimos mejorar la velocidad de actualización del fondo ante un cambio de iluminación, reduciendo el número de falsas detecciones. Por otro lado, en la secuencia $Pets\_005$, en el resultado obtenido mediante la aplicación del método original se aprecia que aparecen algunas detecciones "fantasma" (señaladas por círculos rojos), debidas a la presencia de algunos objetos móviles que se encontraban en esa posición al inicio del análisis de la secuencia y que dieron lugar a la generación de gaussianas que, en la imagen visualizada, todavía tienen una gran influencia. Sin embargo, con el método propuesto, gracias también a la utilización del parámetro C_0, las gaussianas que fueron inicializadas para estos objetos ya no existen, por lo que estos "fantasmas" no aparecen en la detección obtenida.

Para evaluar la calidad de los resultados obtenidos se han utilizado los porcentajes de Recall, Precision, y F, ya definidos en las ecuaciones 3.7, 3.8 y 3.9 de la sección 3.7. En este caso, los parámetros utilizados en estas expresiones hacen referencia a la siguiente información:

- CD: Número de píxeles móviles correctamente detectados.
- ND: Número de píxeles móviles no clasificados como tales.
- FD: Número de falsas detecciones (píxeles del fondo erróneamente etiquetados como móviles).

El resumen de los resultados obtenidos se muestra en la tabla 4.3. Estos resultados muestran que, en el caso de secuencias en las que no se producen cambios permanentes en el fondo, la calidad de los resultados obtenidos con el método propuesto es muy similar a la obtenida con el método original. Sin embargo, en las secuencias en las que el fondo sufre modificaciones prolongadas (Lab_003, Lab_004, Lab_006, $Pets\_001$, $Pets\_004$ y $Pets\_004$), al limitar el tiempo máximo de existencia de las gaussianas no utilizadas, se ha conseguido reducir el número de falsas detecciones (porcentajes de Precision más altos). Por otro lado, en el caso de las secuencias $Wall\_001$ y $Wall\_002$, en las que el fondo es mucho más dinámico que en cualquiera de las otras, se puede observar que el elevado número de falsas detecciones ha dado lugar, independientemente del método utilizado, a valores de Precision muy bajos.

En el caso de la secuencia $Pets\_002$, a lo largo de la cual se suceden varios cambios de iluminación, no hay contenido móvil. Por lo tanto, la calidad de los resultados en esta secuencia ha de ser medida atendiendo a la cantidad de falsas detecciones obtenidas. Aplicando el método clásico se ha obtenido un total de 2785967 falsas detecciones, mientras que con el método propuesto este número se ha reducido a 2160138, lo que equivale a una reducción del $22,46\,\%$ de las falsas detecciones.

4.5. Conclusiones

En este capítulo se ha presentado una estrategia para la detección de objetos móviles en secuencias de vídeo, basada en el popular método de mezcla de gaussianas, capaz de obtener resultados de gran calidad en tiempo real.

Dicha estrategia, a partir del análisis de las variaciones sufridas en cada píxel a lo largo de

su historia reciente, permite seleccionar dinámicamente el número de gaussianas utilizadas por cada píxel en cada instante de tiempo. Así, mediante la asignación dinámica del número de gaussianas en función de las necesidades de cada píxel, reduce muy notablemente el número medio de gaussianas utilizadas en cada imagen, lo cual se traduce en un importante ahorro computacional y de memoria, haciendo factible su uso en aplicaciones que requieren trabajar en tiempo real (vídeo-vigilancia, monitorización, etc.).

Por otro lado, la utilización de un contador asociado a cada gaussiana para determinar en qué momento dejan de ser necesarias permite reducir la dependencia de los resultados con los valores asignados a algunos de los parámetros de los que depende el método original, haciendo más sencilla su utilización y permitiendo utilizar los mismos valores independientemente de las características de la secuencia analizada.

Los resultados obtenidos han mostrado que el método propuesto mejora notablemente la eficiencia computacional del método original. Además, en secuencias en las que el fondo sufre cambios de iluminación o variaciones prolongadas se ha logrado reducir muy apreciablemente la cantidad de píxeles del fondo erróneamente clasificados como móviles.

Capítulo 5

Detección de objetos móviles con técnicas de modelado no paramétrico

Es el porvenir quien debe imperar sobre el pretérito,
y de él recibimos la orden para nuestra conducta
frente a cuanto fue.

José Ortega y Gasset (1883-1955),
filósofo y ensayista español.

Resumen: Para mejorar la calidad de las detecciones en situaciones en las que las variaciones de los píxeles no pueden modelarse haciendo uso de métodos paramétricos como el presentado en el capítulo anterior, en los últimos años se han propuesto algunas estrategias basadas en el modelado no paramétrico. Estas estrategias, mediante la evaluación de funciones locales centradas en las muestras más recientes de cada píxel, proporcionan detecciones de gran calidad. Sin embargo, también tienen algunas limitaciones: suponen un elevado coste de computación y de memoria, requieren de distintas funciones locales en función de las características de las secuencias analizadas y, en situaciones en las que el fondo y los objetos móviles son muy parecidos, no proporcionan detecciones con la suficiente calidad. En este capítulo se describe una eficiente estrategia de detección basada en el modelado no paramétrico del fondo y del primer plano. Para obtener resultados satisfactorios, independientemente de las características de las secuencias analizadas, la estrategia desarrollada estima dinámicamente las características más adecuadas para las funciones locales utilizadas. Además, gracias a una innovadora estrategia basada en un filtro de partículas diseñado para trabajar eficientemente con un número variable de regiones móviles, las posiciones de los objetos móviles previamente detectados se actualizan de imagen a imagen, permitiendo mejorar la calidad del modelado del primer plano y, además, reducir su coste computacional asociado. Adicionalmente, de la aplicación del filtro de partículas se obtiene información a priori que, combinada con los modelos de fondo y primer plano, permite mejorar todavía más la calidad de las detecciones.

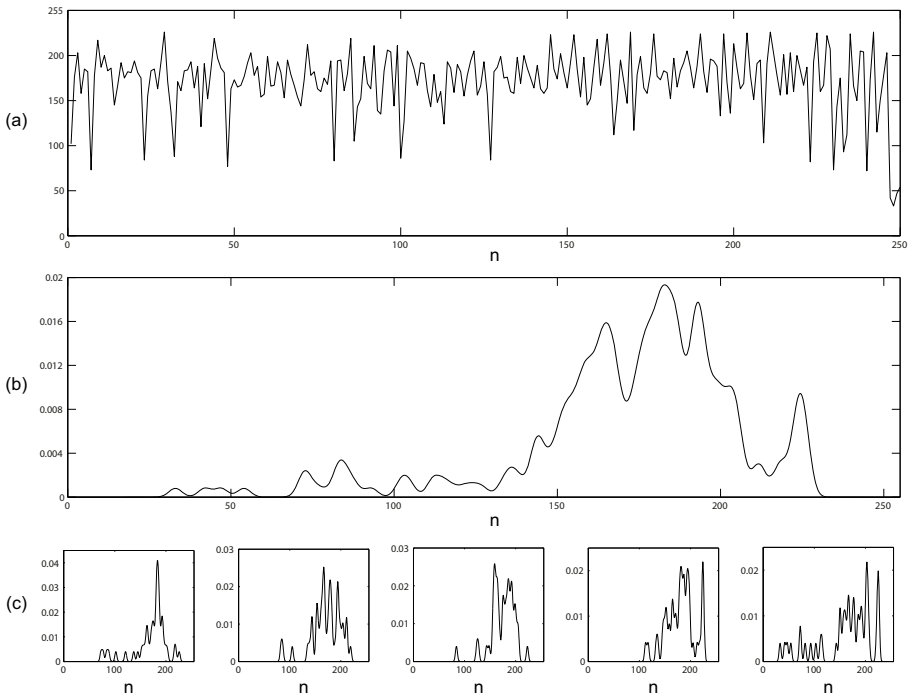


Figura 5.1: Análisis de los valores de un píxel a lo largo de una secuencia de 250 imágenes. (a) Valores de luminancia. (b) Histograma normalizado. (c) Histogramas parciales.

5.1. Introducción

Los métodos de modelado paramétrico, como el presentado en el capítulo 4, proporcionan resultados satisfactorios en un gran número de escenarios con fondos multimodales. Sin embargo, en situaciones en las que las variaciones sufridas por el fondo son muy frecuentes, estos métodos no ofrecen resultados suficientemente buenos (Lu y Hager, 2007). En la figura 5.1 se muestra un ejemplo de esta situación para el caso de un píxel que se ha analizado a lo largo de 250 imágenes consecutivas (10 segundos). La primera de las gráficas de este ejemplo (figura 5.1.a) presenta los valores de luminancia de dicho píxel, mientras que la segunda gráfica (figura 5.1.b) muestra la distribución de probabilidad asociada a dichos valores. Una distribución de este tipo no puede ser adecuadamente estimada haciendo uso de un número pequeño de gaussianas por lo que, en situaciones como la de este ejemplo, el método de mezcla de gaussianas no será capaz de proporcionar resultados satisfactorios. Además, la distribución de los valores de los píxeles puede sufrir cambios muy significativos en periodos cortos de tiempo, tal y como muestran las gráficas de la figura 5.1.c, en las que se han representado las distribuciones asociadas al píxel analizado en intervalos de tiempo de unos 2 segundos. Por lo tanto, es necesario utilizar estrategias capaces de adaptarse rápidamente a estos cambios.

Para conseguir resultados satisfactorios en situaciones en las que las variaciones sufridas por los píxeles no pueden modelarse haciendo uso de métodos paramétricos, en los últi-

mos años se han propuesto algunas estrategias basadas en el modelado no paramétrico (Tavakkoli et al., 2009) (Tanaka et al., 2010) (Ding et al., 2011). Estas estrategias, mediante la superposición de funciones locales centradas en las muestras más recientes de cada píxel, construyen una estimación probabilística para determinar su pertenencia al fondo o al primer plano de la secuencia y, por lo tanto, no asumen que los valores de los píxeles sigan una distribución concreta.

Aunque las estrategias basadas en el modelado no paramétrico consiguen mejorar los resultados en situaciones como la del ejemplo anterior, también tienen algunas limitaciones que deben ser consideradas (Wang et al., 2011). Su principal inconveniente es que, para cada píxel en cada imagen, se debe llevar a cabo la evaluación de una función local centrada en cada una de las muestras utilizadas, lo cual resulta en un elevado coste computacional y de memoria (Elhabian et al., 2008). Otro de sus inconvenientes es la elección de un ancho apropiado para dichas funciones locales (Wan y Wang, 2008). Si son demasiado anchas se reducirá la capacidad del método para representar las variaciones multimodales de los píxeles, mientras que si son demasiado estrechas la estimación resultante será muy ruidosa.

Dentro de estas estrategias se pueden distinguir algunas que, para conseguir mejores resultados en situaciones en las que los objetos móviles presentan características similares a determinadas regiones del fondo, combinan el modelo correspondiente al fondo con un modelado de los objetos móviles (Sheikh y Shah, 2005) (Martel-Brisson y Zaccarin, 2008) (Zhang y Yang, 2008). Sin embargo, estas estrategias, para reducir su elevado coste computacional llevan a cabo algunas simplificaciones (por ejemplo, la utilización de funciones locales con anchos fijos) que condicionan la calidad de sus resultados en función de las características de las secuencias analizadas.

En este capítulo se presenta una estrategia para la detección de objetos móviles, basada en las técnicas de modelado no paramétrico, capaz de conseguir resultados de gran calidad sin que su coste computacional asociado sea excesivamente elevado. Para mejorar los resultados en escenarios en los que los objetos móviles son similares al fondo, haciendo uso de un clasificador bayesiano, se combina un modelo estimado para el fondo con otro estimado para el primer plano. Para reducir el número de falsas detecciones debidas a las posibles vibraciones de las cámaras y a las regiones no estáticas del fondo, ambos modelados se elaboran a partir de la información extraída de imágenes anteriores, dentro de un margen espacial en torno a cada píxel. La calidad de ambos modelados se ha mejorado gracias a la aplicación de estrategias que, de forma muy eficiente, permiten la estimación dinámica de los anchos de las funciones locales utilizadas: con un método basado en un análisis estadístico de las diferencias entre muestras consecutivas para el caso del fondo y, en el caso del primer plano, utilizando un algoritmo basado en *Mean-Shift* (Cheng, 1995) que permite agrupar los píxeles del primer plano con características similares. Además, para conseguir una mejora adicional de los resultados, se ha desarrollado una innovadora estrategia que, actualizando las posiciones de los objetos móviles previamente detectados, permite mejorar la calidad del modelado del primer plano y reducir su coste computacional asociado. Esta actualización se lleva a cabo mediante la aplicación de un filtro de partículas (Doucet et al., 2001) capaz de trabajar con un número variable de regiones móviles y del que, adicionalmente, se obtiene información a priori relativa a las posiciones en las que es más probable la

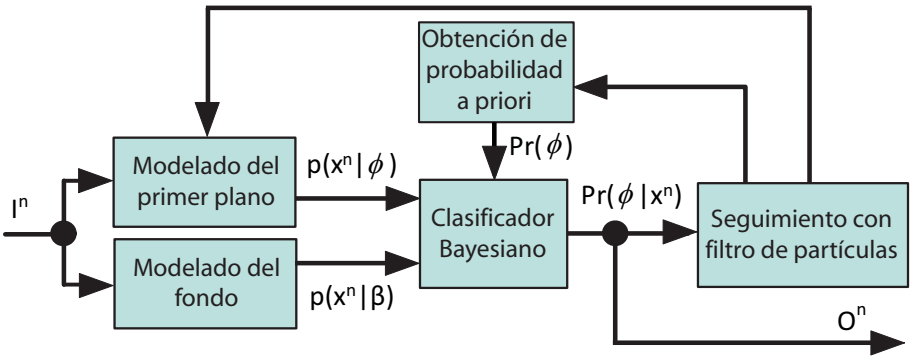


Figura 5.2: Diagrama de bloques del sistema de detección propuesto.

localización de los objetos móviles en el futuro. Para hacer un uso eficiente de esta información se ha propuesto un clasificador bayesiano en el que se combinan los modelos estimados para el fondo y para el primer plano y que permite utilizar cualquier tipo de información a priori.

En primer lugar, en la sección 5.2 se describe la arquitectura de la estrategia desarrollada. A continuación, en la sección 5.3 se describe el modo en el que a partir de la superposición de funciones locales es posible estimar la función densidad de probabilidad un conjunto de muestras aleatorias e independientes. En la sección 5.4 se presentan las estrategias de estimación de las funciones densidad asociadas al fondo y al primer plano, así como el clasificador bayesiano que combina estas funciones y la probabilidad a priori obtenida del filtro de partículas. La estrategia de seguimiento de las regiones móviles y la estrategia de obtención de las probabilidades a priori se describen en la sección 5.5. Seguidamente, en la sección 5.6 se presentan las estrategias para la estimación dinámica de los anchos de las funciones locales utilizadas en los modelados del fondo y del primer plano. Por último, en las secciones 5.7 y 5.8 se presentan los resultados y las conclusiones del capítulo.

5.2. Arquitectura del sistema

La estrategia propuesta, cuyo diagrama de bloques aparece representado en la figura 5.2, contiene dos etapas principales: una basada en la detección de los objetos móviles y otra basada en seguimiento de los objetos detectados.

Para cada nueva imagen, I^n, en el instante n, en primer lugar se construyen los modelos no paramétricos del fondo y del primer plano de las secuencias. Estos modelos son combinados en un clasificador bayesiano del que resulta el conjunto de objetos móviles detectados, O^n. Mediante la aplicación de una estrategia de seguimiento basada en un filtro de partículas se consigue actualizar las coordenadas de las regiones móviles previamente detectadas. Esta información se utiliza para construir el siguiente modelo del primer plano, permitiendo mejorar el resultado de su estimación y, además, dando lugar a una significativa reducción de su tiempo de procesamiento asociado.

Por otro lado, del resultado proporcionado por el filtro de partículas se obtiene infor-

mación a priori relativa la la posición de los objetos móviles en la siguiente imagen. Esta información, combinada con los modelados obtenidos para el fondo y el primer plano, permite obtener una mejora adicional de los resultados.

5.3. Estimación de la densidad de probabilidad mediante la superposición de *kernels*

La estimación de la densidad de probabilidad haciendo uso de funciones locales superpuestas (también conocida como *Kernel Density Estimation* (*KDE*)) fue introducida por *Rosenblatt* en (Rosenblatt, 1956) y por *Parzen* en (Parzen, 1962). Su objetivo consiste en estimar la función densidad de probabilidad de un conjunto de muestras aleatorias e independientes mediante la superposición de funciones locales, comúnmente denominadas funciones núcleo o *kernels*. Este método de estimación, aplicado al campo de la detección de objetos móviles en secuencias de vídeo, permite la obtención de modelos no paramétricos que representan las variaciones de los píxeles en el tiempo.

Considérese un píxel cualquiera, p^n, de la imagen I^n, en el instante temporal n de una secuencia de vídeo. Considérese este píxel definido por un vector de D dimensiones, $x^n \in \mathbb{R}^D$, en el que cada componente es una característica del píxel en cuestión. Sea un conjunto de N muestras D-dimensionales, $\{x^i\}_{i=1}^N$, asociadas a los valores de dicho píxel en las N imágenes previas a la actual. La función densidad de probabilidad (fdp) asociada al valor actual de p^n puede estimarse de forma no paramétrica como (Tavakkoli et al., 2009):

$$\widehat{f}(\mathbf{x}^n) = \frac{1}{N}\sum_{i=1}^{N} K_\Sigma(\mathbf{x}^n - \mathbf{x}^i) \tag{5.1}$$

donde K_Σ es una función definida como:

$$K_\Sigma(\mathbf{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}}} K\left(\frac{\mathbf{x}}{\Sigma^{1/2}}\right) \tag{5.2}$$

y Σ es una matriz simétrica, positiva y con $D \times D$ componentes, que se va a denominar matriz de escala y que determina el ancho del *kernel* K (Turlach, 1993).

Normalmente, para facilitar su utilización, a los *kernels* utilizados en la estimación de funciones densidad de probabilidad se les imponen algunas condiciones (Elgammal et al., 2002) (Sheikh y Shah, 2005) como, por ejemplo:

- Que su integral esté normalizada: $\int K(\mathbf{x})d\mathbf{x} = 1$
- Que sean simétricos: $K(\mathbf{x}) = K(-\mathbf{x})$
- Que no sean negativos: $K(\mathbf{x}) \geq 0$
- Que tengan media nula: $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0$
- Que su covarianza sea proporcional a la matriz identidad: $\int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} \propto I$

En la tabla 5.1 se muestra un resumen con algunos de los *kernels* más populares (Turlach, 1993), para el caso de $D = 1$. Cualquiera de las funciones representadas en dicha tabla puede ser utilizada como *kernel* para la estimación de una función densidad de probabilidad, y

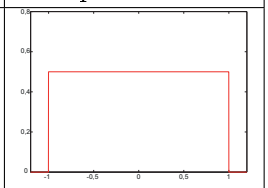
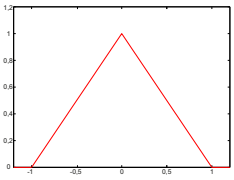
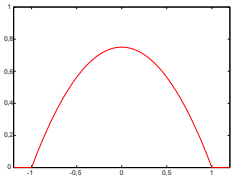
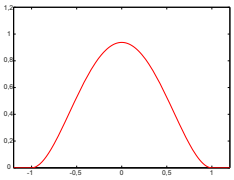
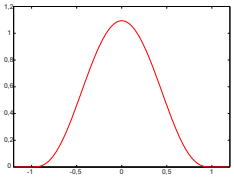
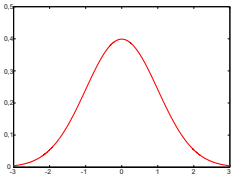
kernel	$K(\mathbf{x})$	Representación
Uniforme	$\begin{cases} \frac{1}{2} & \text{si } \lvert\mathbf{x}\rvert \leq 1 \\ 0 & \text{si } \lvert\mathbf{x}\rvert > 1 \end{cases}$	
Triangular	$\begin{cases} 1 - \lvert\mathbf{x}\rvert & \text{si } \lvert\mathbf{x}\rvert \leq 1 \\ 0 & \text{si } \lvert\mathbf{x}\rvert > 1 \end{cases}$	
Epanechnikov	$\begin{cases} \frac{3}{4}\left(1 - \mathbf{x}^2\right) & \text{si } \lvert\mathbf{x}\rvert \leq 1 \\ 0 & \text{si } \lvert\mathbf{x}\rvert > 1 \end{cases}$	
Quartic	$\begin{cases} \frac{15}{16}\left(1 - \mathbf{x}^2\right)^2 & \text{si } \lvert\mathbf{x}\rvert \leq 1 \\ 0 & \text{si } \lvert\mathbf{x}\rvert > 1 \end{cases}$	
Triweight	$\begin{cases} \frac{35}{32}\left(1 - \mathbf{x}^2\right)^3 & \text{si } \lvert\mathbf{x}\rvert \leq 1 \\ 0 & \text{si } \lvert\mathbf{x}\rvert > 1 \end{cases}$	
Gaussiano	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mathbf{x}^2\right)$	

Tabla 5.1: Ejemplos de *kernels* comúnmente utilizados.

cada una de ellas tiene sus ventajas y sus inconvenientes (Wand y Jones, 1995). Nosotros, debido a sus características de continuidad y diferenciabilidad, y a sus propiedades locales (Elgammal et al., 2002), hemos decidido utilizar *kernels* gaussianos.

La estimación de densidades de probabilidad con *kernels* gaussianos se puede ver como una generalización del método de mezcla de gaussianas presentado en el capítulo 4, en la que cada muestra de las N utilizadas es considerada como una distribución gaussiana, $N(0, \Sigma)$, por sí misma. Por lo tanto, el modelado no paramétrico con *kernels* permite una estimación mucho más precisa de la función de densidad de probabilidad asociada a los datos analizados (Elgammal et al., 2000), en la que se evitan los errores comunes de

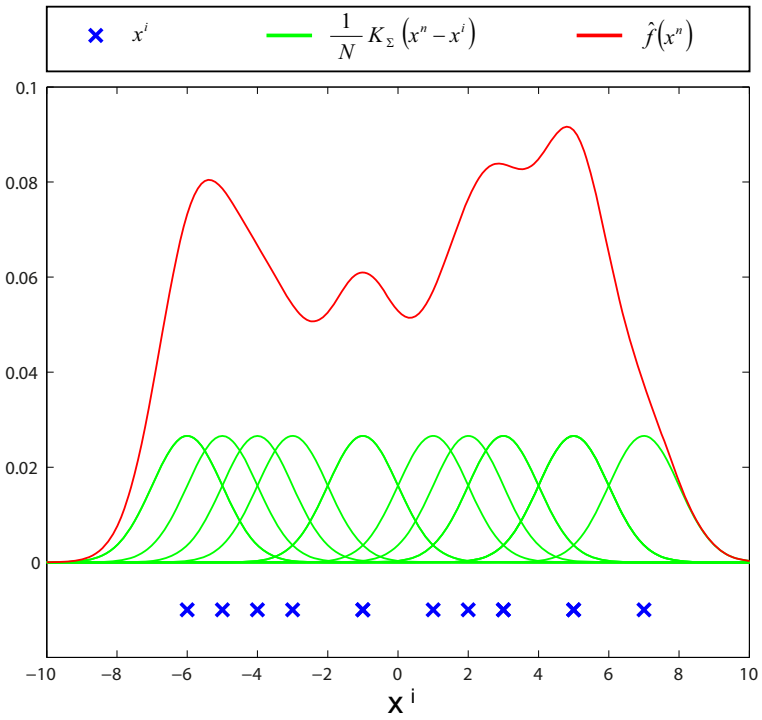


Figura 5.3: Estimación de una función densidad de probabilidad utilizando *kernels* gaussianos.

los métodos paramétricos en los que, normalmente, para obtener estimaciones precisas se requiere hacer uso de grandes cantidades de datos. Otra ventaja de las estrategias de modelado no paramétrico es su capacidad para ignorar rápidamente el pasado y centrarse en el presente, ya que la información que utilizan es siempre la más reciente, lo cual mejora la calidad de los resultados en situaciones en las que el fondo de las secuencias sufre cambios frecuentes.

En la figura 5.3 se muestra un ejemplo en el que se ha estimado la función densidad de probabilidad (representada en rojo) de un conjunto de muestras (representadas con cruces azules sobre el eje de abscisas), utilizando *kernels* gaussianos centrados en dichas muestras (representados con líneas verdes).

5.3.1. Estimación del ancho de las funciones locales

La elección de una matriz de escala, Σ, que determine un ancho adecuado para los *kernels* utilizados es fundamental para obtener unos resultados satisfactorios (Wan y Wang, 2008). Por este motivo, en la literatura reciente es posible localizar un amplio número de trabajos que proponen distintas alternativas para su estimación (Martel-Brisson y Zaccarin, 2008) (Liao et al., 2010).

Por un lado, los *kernels* con un ancho pequeño son más apropiados en entornos con una densidad elevada de datos. Sin embargo, si son demasiado estrechos, la función de densidad

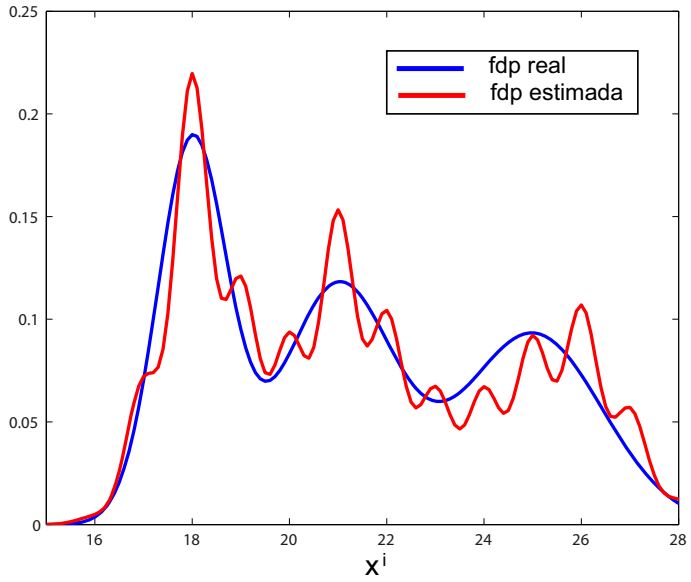


Figura 5.4: Estimación de una función densidad de probabilidad con *kernels* estrechos.

a la que dan lugar puede llegar a ser demasiado ruidosa. Por otro lado, si la densidad de datos es baja son más adecuados los *kernels* con un ancho mayor, pero si son excesivamente anchos pueden dar lugar a estimaciones poco precisas.

Para mostrar la importancia de la elección de *kernels* con un ancho adecuado se ha analizado el resultado de la estimación de una función densidad de probabilidad definida como:

$$f(\mathbf{x}) = \frac{1}{3}N\left(18, \frac{5}{7}\right) + \frac{1}{3}N\left(21, \frac{8}{7}\right) + \frac{1}{3}N\left(25, \frac{10}{7}\right) \tag{5.3}$$

utilizando 600 observaciones y aplicando *kernels* gaussianos con distinto ancho. En la figura 5.4 se ha representado el resultado obtenido con $\Sigma = 0{,}35$, mientras que en la figura 5.5 se ha representado el caso correspondiente a la utilización de *kernels* con $\Sigma = 0{,}6$. En el primer caso, al haberse utilizado *kernels* demasiado estrechos, la estimación obtenida resulta excesivamente ruidosa. Por el contrario, en el segundo caso, al haberse utilizado *kernels* demasiado anchos, la estimación obtenida es poco precisa y se aprecia la pérdida de capacidad del método para detectar todos los modos de la función densidad de probabilidad estimada.

Teóricamente, la matriz Σ óptima se puede conseguir minimizando el error cuadrático medio entre la función densidad de probabilidad estimada, $\widehat{f}_\Sigma(\mathbf{x})$, y la densidad real, $f(\mathbf{x})$ (Sheikh y Shah, 2005):

$$MSE\left(\widehat{f}\mathbf{x})\right) = E\left(\left[\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2\right) \tag{5.4}$$

Dado que este error cuadrático medio depende de la función de densidad real, la cual

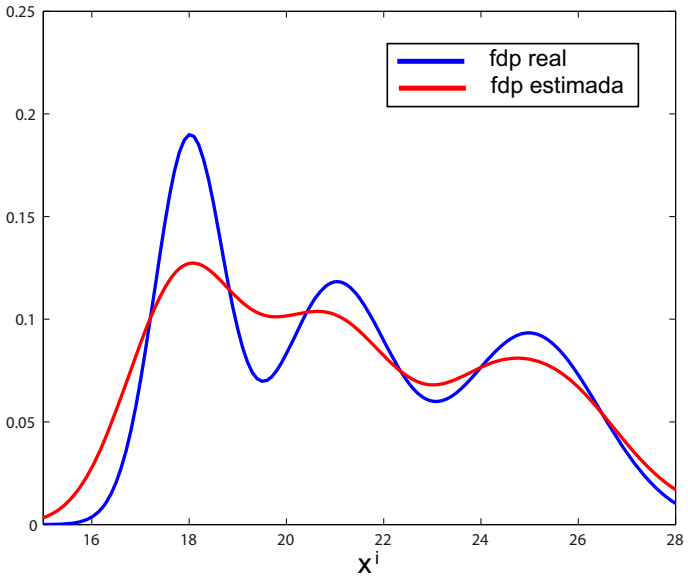


Figura 5.5: Estimación de una función densidad de probabilidad con *kernels* anchos.

es desconocida, en los últimos años se han publicado numerosas propuestas heurísticas que, en función de los datos utilizados, estiman dinámicamente las matrices de escala más adecuadas. Estas estrategias muestran dos formulaciones. La primera se denomina *balloon estimator* y se caracteriza por asignar distinto ancho a cada *kernel* en función del dato que se está estimando:

$$\widehat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|\Sigma(\mathbf{x})|^{\frac{1}{2}}} K\left(\frac{\mathbf{x}-\mathbf{x}^i}{(\Sigma(\mathbf{x}))^{1/2}}\right) \tag{5.5}$$

donde $\Sigma(\mathbf{x})$ es la matriz de escala que determina el ancho del *kernel* asociado a $\mathbf{x}$.

La segunda formulación se conoce como *sample-point estimator* y se caracteriza por variar el ancho de los *kernels* en función de cada una de las muestras de referencia utilizadas:

$$\widehat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{|\Sigma(\mathbf{x}^i)|^{\frac{1}{2}}} K\left(\frac{\mathbf{x}-\mathbf{x}^i}{(\Sigma(\mathbf{x}^i))^{1/2}}\right) \tag{5.6}$$

donde $\Sigma(\mathbf{x}^i)$ es la matriz de escala que determina el ancho del *kernel* asociado a la muestra $\mathbf{x}^i$.

En la literatura, dependiendo de los propósitos y de las características de la estrategia aplicada, se pueden encontrar distintas alternativas para estimación de las matrices de escala. Algunas parametrizan completamente dichas matrices, con lo que consiguen mejorar la calidad de los resultados aunque a consta de un notable aumento de computación y de memoria (Tavakkoli et al., 2009). Otras utilizan matrices diagonales, logrando un buen compromiso entre calidad y velocidad de procesamiento (Martel-Brisson y Zaccarin, 2008).

Por último, las que pretenden alcanzar grandes velocidades de procesamiento no estiman el ancho de los *kernels* y utilizan matrices diagonales con valores fijos (Zhang y Yang, 2008). En la sección 5.6 se exponen con más detalle las virtudes y los inconvenientes de cada una de estas estrategias de estimación.

5.4. KDE aplicada a la detección de objetos móviles

En esta sección se describen las estrategias desarrolladas para llevar a cabo el modelado no paramétrico del fondo y del primer plano. Estas estrategias, para explotar las dependencias espaciales de los píxeles, tienen en cuenta no sólo la información de imágenes previas en una misma posición espacial, sino también la información de todos los píxeles dentro de un vecindario. Además, en esta sección también se describe el clasificador bayesiano en el que se combina la información resultante de dichos modelados y del que se obtiene la probabilidad de cada píxel de formar parte del fondo o del primer plano de las secuencias.

5.4.1. Modelado no paramétrico del fondo

En la estrategia descrita en el capítulo 4, la detección de objetos móviles se llevaba a cabo haciendo uso de D características de apariencia de los píxeles. Sin embargo, en los algoritmos descritos en este capítulo, también se va a hacer uso de las 2 componentes espaciales que determinan la posición de los píxeles dentro de las imágenes. Por lo tanto, si se considera un píxel cualquiera, p^n, en el instante n, dicho píxel estará definido por un vector de $D+2$ características, $\mathbf{x}^n = ((\mathbf{c}^n)^T, (\mathbf{s}^n)^T)^T \in \mathbb{R}^{D+2}$, a su vez compuesto por dos vectores, $\mathbf{c}^n \in \mathbb{R}^D$ y $\mathbf{s}^n \in \mathbb{R}^2$. Las D primeras características de $\mathbf{x}^n$ son las que conforman el vector $\mathbf{c}^n$ y hacen referencia a la información de apariencia del píxel (su color, su gradiente, etc.), por lo que se han denominado características de apariencia. Las 2 últimas características de $\mathbf{x}^n$ conforman el vector $\mathbf{s}^n$ y hacen referencia a las coordenadas espaciales de los píxeles dentro de las imágenes, por lo que se han denominado características espaciales o de posición.

Considérese también que el píxel p^n tiene asociado un conjunto de N_β muestras de referencia, $\{\mathbf{x}_\beta^i = ((\mathbf{c}_\beta^i)^T, (\mathbf{s}_\beta^i)^T)^T\}_{i=1}^{N_\beta}$, extraídas de los píxeles pertenecientes a las N_{β_N} imágenes previas a la actual, en un entorno espacial definido por un ancho de banda vertical, σ_{β_H}, y otro ancho de banda horizontal, σ_{β_W}, en torno a las coordenadas de p^n.

De acuerdo con el modelo de estimación expuesto en la sección 5.3 y haciendo uso de *kernels* gaussianos, la verosimilitud del píxel p^n de pertenecer al fondo de la secuencia, β, puede ser estimada de forma no paramétrica como:

$$\widehat{f}_\beta(\mathbf{x}^n) = p(\mathbf{x}^n|\beta) = \frac{1}{N_\beta}\sum_{i=1}^{N_\beta} \frac{1}{(2\pi)^{\frac{D+2}{2}}|\Sigma_\beta|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}^n - \mathbf{x}_\beta^i)^T\Sigma_\beta^{-1}(\mathbf{x}^n - \mathbf{x}_\beta^i)\right) \qquad (5.7)$$

donde Σ_β es la matriz escala, de $(D+2)\times(D+2)$ componentes, que determina el ancho de los *kernels*.

En la figura 5.6 se ha representado un ejemplo en el que se muestra el conjunto de píxeles de referencia que serían utilizados para obtener la verosimilitud de un píxel de la

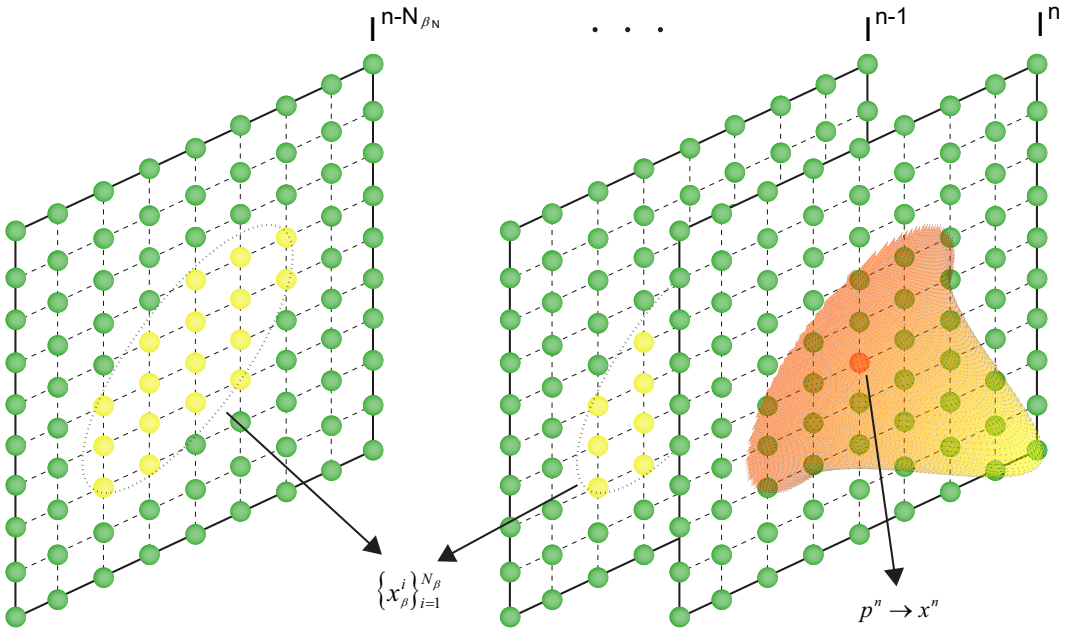


Figura 5.6: Píxeles utilizados para estimar la función densidad de probabilidad del fondo.

imagen actual. Sobre dicho píxel, representado en color rojo sobre la imagen I^n de la figura, se ha dibujado una gaussiana bidimensional que representa la influencia de las componentes espaciales de los *kernels* gaussianos utilizados. Los píxeles de referencia utilizados, representados en amarillo, serán todos los que, perteneciendo a las N_{β_N} imágenes previas a la actual, verifiquen que su distancia de *Mahalanobis* con respecto al píxel p^n es menor que 3:

$$\begin{pmatrix} \Delta h & \Delta w \end{pmatrix} \begin{pmatrix} \sigma^2_{\beta_H} & \sigma_{\beta_H}\sigma_{\beta_W} \\ \sigma_{\beta_H}\sigma_{\beta_W} & \sigma^2_{\beta_W} \end{pmatrix}^{-1} \begin{pmatrix} \Delta h \\ \Delta w \end{pmatrix} \leq 3^2 \tag{5.8}$$

siendo $(\Delta h, \Delta w)$ el vector que representa la distancia espacial, en filas y columnas, entre cada muestra de referencia y el píxel p^n. De esta forma se garantiza que se están utilizando todos los píxeles contenidos en aproximadamente el 99 % de la gaussiana bidimensional (ver apéndice B) y se evita realizar operaciones sobre píxeles más lejanos que apenas influyen en el resultado de la estimación. Así, al reducirse el número de operaciones, se consigue un importante ahorro computacional en el modelado del fondo.

5.4.2. Modelado no paramétrico del primer plano

En ocasiones las regiones móviles pueden presentar características similares a determinadas regiones del fondo de las secuencias. En estos casos la detección de objetos móviles basada únicamente en el modelado del fondo no es suficiente para discriminar correctamente entre el fondo y los objetos móviles (Zhang y Yang, 2008). Como solución a este problema se hace necesaria la utilización de un modelado de los objetos móviles que, combinado con el modelo

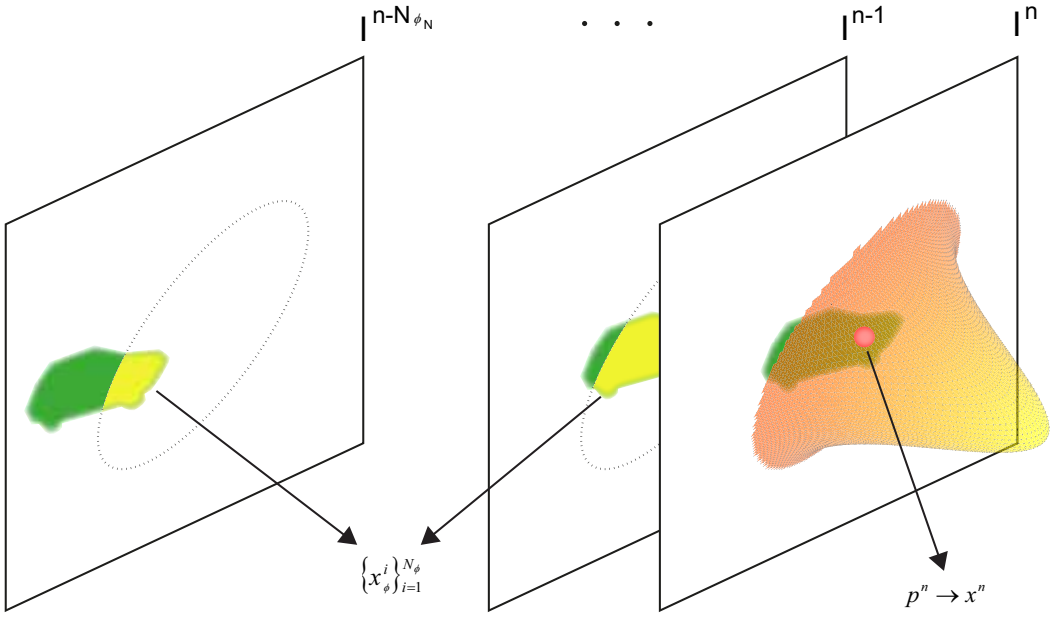


Figura 5.7: Píxeles utilizados para estimar la función densidad de probabilidad del primer plano.

obtenido para el fondo, permita mejorar la calidad de los resultados en estas situaciones.

Considérese un píxel cualquiera, p^n, en la imagen I^n. Considérese dicho píxel definido por el vector de características x^n (descrito al comienzo de la sección 5.4.1). En principio, la probabilidad de que ese píxel sea parte del primer plano es la misma que la de cualquier otro píxel, independientemente de su posición dentro de la imagen y de sus características de apariencia. Por lo tanto, dicha probabilidad puede ser representada con una función densidad uniforme. Sin embargo, si en imágenes previas a la actual se han detectado regiones móviles en su entorno, la probabilidad de observar una región del primer plano con características de apariencia y de posición similares a las de estas regiones se habrá incrementado. Por lo tanto, la verosimilitud del píxel p^n de pertenecer al primer plano de la secuencia, ϕ, puede construirse a partir de la mezcla de la función uniforme previamente mencionada y una función de densidad construida con *kernels* gaussianos (Sheikh y Shah, 2005):

$$\widehat{f}_\phi(\mathbf{x}^n) = p(\mathbf{x}^n|\phi) = \alpha\gamma + \frac{(1-\alpha)}{N_\phi}\sum_{i=1}^{N_\phi}\frac{1}{(2\pi)^{\frac{D+2}{2}}|\Sigma_\phi|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}^n - \mathbf{x}_\phi^i)^T\Sigma_\phi^{-1}(\mathbf{x}^n - \mathbf{x}_\phi^i)\right) \quad (5.9)$$

en la que Σ_ϕ es la matriz de escala que determina el ancho de los *kernels* gaussianos, $\alpha \in [0,1]$ es el factor de mezcla de las dos funciones (típicamente, $\alpha \ll 1$), γ es un valor de densidad constante que se corresponde con una variable aleatoria uniforme en el espacio de características utilizadas y $\{\mathbf{x}_\phi^i = ((\mathbf{c}_\phi^i)^T, (\mathbf{s}_\phi^i)^T)^T\}_{i=1}^{N_\phi}$ es el conjunto de muestras de referencia, extraídas de los píxeles clasificados como parte del primer plano a lo largo de

las N_{ϕ_N} imágenes previas, en un entorno espacial definido por un ancho de banda vertical, σ_{ϕ_H}, y otro ancho de banda horizontal, σ_{ϕ_W}, en torno al píxel p^n.

En la figura 5.7 se muestra un ejemplo que relaciona a un píxel cualquiera de la imagen actual (representado en color rojo), con el conjunto de muestras de referencia que serían utilizadas para obtener la verosimilitud de su pertenencia al primer plano de una secuencia. Estas muestras (representadas en color amarillo) son las asociadas a todos los píxeles de las N_{ϕ_N} imágenes previas a la actual que, al igual que en el ejemplo descrito previamente para el modelado del fondo, representan aproximadamente el 99 % de las gaussianas espaciales utilizadas en el modelado y, por lo tanto (ver apéndice B), han de verificar que:

$$\begin{pmatrix} \Delta h & \Delta w \end{pmatrix} \begin{pmatrix} \sigma^2_{\phi_H} & \sigma_{\phi_H}\sigma_{\phi_W} \\ \sigma_{\phi_H}\sigma_{\phi_W} & \sigma^2_{\phi_W} \end{pmatrix}^{-1} \begin{pmatrix} \Delta h \\ \Delta w \end{pmatrix} \leq 3^2 \tag{5.10}$$

5.4.3. Clasificador Bayesiano

Una vez obtenidos los modelos para el fondo y para el primer plano, mediante la utilización de un clasificador bayesiano es posible obtener, para cada píxel, un valor de probabilidad de su pertenencia a una de las dos clases.

Partiendo del teorema de *Bayes* (Bayes y Price, 1763), la probabilidad de que el píxel p^n pertenezca a un objeto móvil se puede calcular como:

$$Pr(\phi|\mathbf{x}^n) = \frac{Pr(\phi)p(\mathbf{x}^n|\phi)}{Pr(\phi)p(\mathbf{x}^n|\phi) + Pr(\beta)p(\mathbf{x}^n|\beta)} \tag{5.11}$$

donde $Pr(\phi)$ y $Pr(\beta)$ son las probabilidades a priori del primer plano y del fondo.

En principio, si no se dispone de ninguna información adicional a la proporcionada por los modelos estimados para el fondo y para el primer plano, las probabilidades a priori de ambas clases deberían ser iguales, $Pr(\phi) = Pr(\beta) = \frac{1}{2}$ (Landabaso y Pardas, 2008) (Tran et al., 2009). Sin embargo, normalmente se dispone de información adicional que podría dar pistas sobre las zonas de la imagen en las que es más probable la presencia de los objetos móviles como, por ejemplo, la posición de dichos objetos en instantes anteriores, su trayectoria de desplazamiento, o su velocidad. En el modelado del primer plano, tal y como se ha descrito en la sección 5.4.2, se está teniendo en cuenta información a priori relativa a la localización de los objetos móviles en instantes anteriores. Sin embargo, haciendo uso de un clasificador como el descrito en la ecuación 5.11, en el que la probabilidad a priori es una constante de igual valor en cualquier posición espacial, no es posible incluir la probabilidad a priori proporcionada por las trayectorias que siguen los objetos móviles, o por sus velocidades de desplazamiento.

Nosotros, para poder hacer uso de esta información o de cualquier otra de la que se disponga, proponemos utilizar un clasificador alternativo que permite hacer uso de una probabilidad a priori dependiente de la posición espacial en la que se encuentre el píxel analizado y que, además, puede ser estimada a partir de la combinación de distintas fuentes de información (posición previa de los objetos móviles, velocidad de los mismos, etc.).

El clasificador propuesto obtiene la probabilidad de primer plano, dado el píxel p^n, a partir de la siguiente factorización de $Pr(\phi|\mathbf{x}^n)$:

$$Pr(\phi|\mathbf{x}^n) = Pr(\phi|\mathbf{c}^n, \mathbf{s}^n) = \frac{p(\mathbf{c}^n, \phi|\mathbf{s}^n)}{p(\mathbf{c}^n|\mathbf{s}^n)} = \frac{Pr(\phi|\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n, \phi)}{p(\mathbf{c}^n, \phi|\mathbf{s}^n) + p(\mathbf{c}^n, \beta|\mathbf{s}^n)} = \\ = \frac{Pr(\phi|\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n, \phi)}{Pr(\phi|\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n, \phi) + Pr(\beta|\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n, \beta)} \tag{5.12}$$

En esta expresión, $Pr(\phi|\mathbf{s}^n)$ y $Pr(\beta|\mathbf{s}^n) = 1 - Pr(\phi|\mathbf{s}^n)$ son las probabilidades a priori del primer plano y del fondo (ambas dependientes del espacio), y $p(\mathbf{c}^n|\mathbf{s}^n, \phi)$ y $p(\mathbf{c}^n|\mathbf{s}^n, \beta)$ son las funciones densidad de probabilidad del primer plano y del fondo, condicionadas en el espacio:

$$p(\mathbf{c}^n|\mathbf{s}^n, \phi) = \frac{p(\mathbf{x}^n|\phi)}{p(\mathbf{s}^n|\phi)} \tag{5.13}$$

$$p(\mathbf{c}^n|\mathbf{s}^n, \beta) = \frac{p(\mathbf{x}^n|\beta)}{p(\mathbf{s}^n|\beta)} \tag{5.14}$$

donde $p(\mathbf{s}^n|\phi)$ y $p(\mathbf{s}^n|\beta)$ son las correspondientes probabilidades marginales en el conjunto de características de apariencia de los píxeles. Dichas probabilidades se calculan como:

$$p(\mathbf{s}^n|\phi) = \alpha\gamma' + \frac{(1-\alpha)}{N_\phi 2\pi} \sum_{i=1}^{N_\phi} \frac{1}{|\Sigma_{\phi,\mathbf{s}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{s}^n - \mathbf{s}_\phi^i)^T \Sigma_{\phi,\mathbf{s}}^{-1}(\mathbf{s}^n - \mathbf{s}_\phi^i)\right) \tag{5.15}$$

$$p(\mathbf{s}^n|\beta) = \frac{1}{N_\beta 2\pi} \sum_{i=1}^{N_\beta} \frac{1}{|\Sigma_{\beta,\mathbf{s}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{s}^n - \mathbf{s}_\beta^i)^T \Sigma_{\beta,\mathbf{s}}^{-1}(\mathbf{s}^n - \mathbf{s}_\beta^i)\right) \tag{5.16}$$

donde $\Sigma_{\phi,\mathbf{s}}$ y $\Sigma_{\beta,\mathbf{s}}$ son las matrices de 2×2 dimensiones que determinan el ancho de los *kernels* en sus componentes espaciales, $\alpha \in [0, 1]$ es el mismo factor de mezcla utilizado en la ecuación 5.9 y γ' es un valor de densidad constante que se corresponde con una variable aleatoria uniforme en el espacio de las características espaciales de los píxeles. Dado que, como se verá más adelante, el ancho de los *kernels* se va a determinar mediante matrices de escala diagonales, estas matrices se pueden extraer directamente de Σ_ϕ y Σ_β. Por lo tanto, las expresiones descritas en las ecuaciones 5.13 y 5.14 pueden obtenerse, sin aumentar el coste computacional del sistema, al mismo tiempo que se estiman las funciones densidad de probabilidad del primer plano (ecuación 5.9) y del fondo (ecuación 5.7).

Como se verá en la sección 5.5.2, en la que se describe el modo de estimación de las probabilidades a priori, la utilización del clasificador alternativo propuesto hace posible combinar los modelos estimados para el fondo y el primer plano, a la vez que se utiliza información a priori proporcionada por posiciones, trayectorias y velocidades de los objetos móviles previamente detectados.

5.5. Seguimiento de los objetos móviles detectados

Una de las principales diferencias entre otras estrategias basadas en el modelado no paramétrico y la aquí propuesta es la etapa correspondiente a la actualización espacial de las regiones móviles previamente detectadas. Esta actualización, aplicada sobre las muestras de referencia utilizadas para modelar el primer plano, permite mejorar la calidad de los resultados a la vez que reduce la carga computacional asociada a dicho modelado.

Para llevar a cabo esta actualización se ha utilizado una estrategia que hace uso de un filtro de partículas (Doucet et al., 2001) que permite seguir múltiples regiones móviles, gestionando su aparición y desaparición sin realizar ningún tipo de asunción. Además, como resultado de la aplicación de este filtro se obtiene información relativa a las localizaciones en las que se espera que se encuentren los objetos móviles en el futuro. De esta información se obtiene la probabilidad a priori utilizada en el clasificador bayesiano descrito en la sección 5.4.3.

A continuación, en las secciones 5.5.1 y 5.5.2 se describe, respectivamente, el proceso de actualización de las posiciones de las regiones móviles y el modo de obtención de las probabilidades a priori. El filtro de partículas utilizado se describe detalladamente en el apéndice C.

5.5.1. Actualización de las posiciones de las regiones del primer plano

Considérese una región del primer plano, correspondiente a un objeto móvil detectado en la imagen I^n. Normalmente, la información de los píxeles de esta región (color, gradiente, etc.) mantendrá valores similares a lo largo del tiempo. Por lo tanto, idealmente, toda esa información debería ser utilizada para estimar la función densidad de probabilidad del primer plano en las siguientes imágenes. Si se tiene en cuenta que la posición espacial de estos datos estará cambiando de cada imagen a la siguiente, para que el modelado del primer plano utilice la información de esta región, el ancho de los *kernels* en sus componentes espaciales deberá ser el suficiente como para cubrir los desplazamientos de esta región a lo largo de las siguientes imágenes. Sin embargo, cuanto mayor sea este ancho espacial, el coste computacional asociado al modelado del primer plano será mayor, ya que cada píxel de la imagen actual deberá ser comparado con un mayor número de muestras. Además, si el ancho de banda espacial utilizado es demasiado grande, otras regiones de las imágenes (tanto estáticas como móviles) con información similar a la de esta región móvil pueden ser inadecuadamente afectadas por ella, empeorando la calidad del modelado.

En la figura 5.8.a se muestra un ejemplo de esta situación para el caso de una secuencia con tres objetos móviles. En la imagen I^{n_3} de este ejemplo se puede observar un objeto móvil, representado en color verde, del que se quiere estimar su función densidad de probabilidad de ser parte del primer plano de la secuencia. Si para llevar a cabo esta estimación se desea utilizar las muestras correspondientes a este mismo objeto en las imágenes anteriores (muestras representadas en amarillo sobre las imágenes I^{n_1} e I^{n_2}), será necesario utilizar *kernels* con un ancho espacial suficientemente grande (representado mediante una elipse sobre las tres imágenes del ejemplo). La utilización de estos anchos espaciales, además de suponer una gran carga computacional (debido al elevado número de muestras que han de

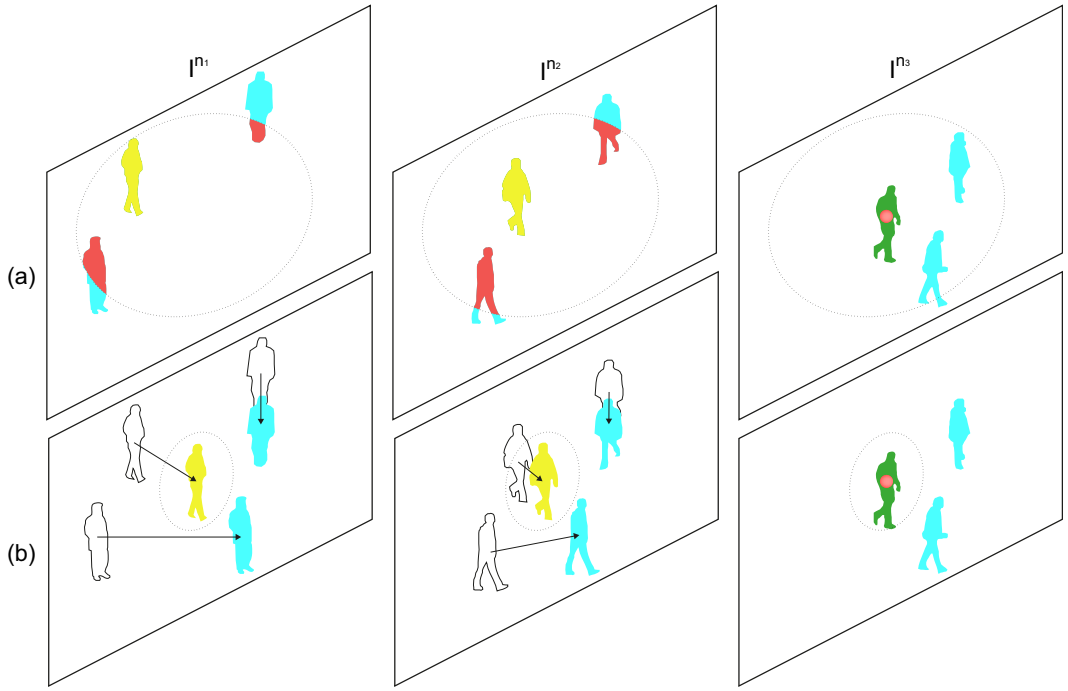


Figura 5.8: Ancho de banda espacial en el modelado del primer plano. (a) Método original, sin actualizar las posiciones de las regiones móviles. (b) Método propuesto, actualizando las posiciones espaciales.

ser comparadas con cada píxel de la imagen actual), hace que se estén teniendo en cuenta muestras pertenecientes a píxeles de otros objetos (representadas en color rojo), lo cual reduce la calidad del modelado.

Para evitar estas limitaciones la estrategia propuesta actualiza las posiciones de las regiones móviles previamente detectadas. De este modo, las muestras utilizadas para modelar a cada objeto móvil se encontrarán en posiciones espaciales similares a las de dicho objeto en el instante actual, permitiendo hacer uso de *kernels* con menor ancho espacial e incrementando la influencia de estas muestras sobre la región a la que pertenecen. Además, al utilizarse *kernels* más estrechos en el espacio, se estará evitando la influencia negativa de otras regiones con características similares.

En la figura 5.8.b se muestra un ejemplo de la aplicación de estas actualizaciones sobre la misma secuencia utilizada en el ejemplo anterior. En este caso, los desplazamientos efectuados sobre las regiones móviles previamente detectadas (indicados mediante flechas) permiten la utilización de un ancho espacial mucho menor (representado mediante una elipse sobre las tres imágenes del ejemplo) y, por lo tanto, se reduce tanto la influencia de unos objetos móviles sobre otros como en número de píxeles que han de ser evaluados. Además, este desplazamiento incrementa la influencia de las muestras del primer plano sobre los objetos móviles a los que pertenecen, permitiendo reducir el número de imágenes de referencia sin que se reduzca la calidad del modelado. Esta reducción del ancho de banda espacial

y del número de imágenes de referencia se traduce en una importante reducción del coste computacional requerido por la etapa de modelado del primer plano.

5.5.2. Probabilidades a priori

En esta sección se describe el modo de estimación de las probabilidades a priori que se utilizan en el clasificador bayesiano alternativo propuesto en la sección 5.4.3. Para tal propósito se ha tratado de combinar lo más eficientemente posible toda la información de la que se dispone acerca de dónde es más o menos probable la presencia de los objetos móviles en cada instante: la posición de los objetos móviles en imágenes previas y las trayectorias y velocidades de desplazamiento de dichos objetos. La información correspondiente a la posición de los objetos móviles en instantes anteriores al actual se obtiene a partir de las probabilidades marginales descritas en las ecuaciones 5.15 y 5.16 de la sección 5.4.3, mientras que la información relativa a sus trayectorias y velocidades se obtiene de la predicción resultante de la aplicación del filtro de partículas utilizado para seguir dichos objetos móviles.

En el clasificador descrito en la ecuación 5.11, al aplicarse directamente los modelos estimados para el fondo y el primer plano, se está teniendo en cuenta implícitamente la posición de los objetos en instantes previos al actual. De esta forma, a igualdad de características de apariencia, la densidad de probabilidad de las muestras de referencia será mayor cuanto menor sea su distancia espacial al píxel bajo análisis. Sin embargo, en el clasificador alternativo propuesto, al aplicarse las funciones densidad de probabilidad condicionadas en el espacio, la posición espacial de las muestras de referencia no está siendo tenida en cuenta. Por lo tanto, si se desea hacer uso de esta información espacial, las probabilidades a priori utilizadas en el clasificador alternativo deberán contemplar las probabilidades marginales descritas en las ecuaciones 5.15 y 5.16.

Por otro lado, como resultado de la aplicación del filtro de partículas sobre la imagen I^{n-1}, en el instante correspondiente a la imagen I^n se dispone de un conjunto de partículas predichas. Dichas partículas estarán distribuidas de forma que su cantidad será mayor en las regiones de la imagen en las que, en función de la trayectoria y la velocidad de desplazamiento de los objetos móviles previamente detectados, es más probable la presencia de regiones pertenecientes al primer plano. Por lo tanto, si se desea tener en cuenta estas trayectorias y velocidades, la información proporcionada por las partículas predichas también deberá ser utilizada para estimar la probabilidad a priori.

Para hacer uso tanto de las probabilidades marginales como de la predicción obtenida con el filtro de partículas, se propone utilizar unas probabilidades a priori definidas, para cada píxel p^n, como:

$$Pr(\phi|\mathbf{s}^n) = \frac{Pr_\phi(\mathbf{s}^n)p(\mathbf{s}^n|\phi)}{Pr_\phi(\mathbf{s}^n)p(\mathbf{s}^n|\phi) + Pr_\beta(\mathbf{s}^n)p(\mathbf{s}^n|\beta)} \tag{5.17}$$

$$Pr(\beta|\mathbf{s}^n) = \frac{Pr_\beta(\mathbf{s}^n)p(\mathbf{s}^n|\beta)}{Pr_\phi(\mathbf{s}^n)p(\mathbf{s}^n|\phi) + Pr_\beta(\mathbf{s}^n)p(\mathbf{s}^n|\beta)} \tag{5.18}$$

donde $Pr_\phi(\mathbf{s}^n)$ es la probabilidad obtenida de la predicción proporcionada por el filtro de partículas y $Pr_\beta(\mathbf{s}^n) = 1 - Pr_\phi(\mathbf{s}^n)$ es su probabilidad complementaria. De este modo, el

clasificador definido en la ecuación 5.12 puede formularse como:

$$Pr(\phi|\mathbf{x}^n) = \frac{Pr_\phi(\mathbf{s}^n)p(\mathbf{s}^n|\phi)p(\mathbf{c}^n|\mathbf{s}^n,\phi)}{Pr_\phi(\mathbf{s}^n)p(\mathbf{s}^n|\phi)p(\mathbf{c}^n|\mathbf{s}^n,\phi) + Pr_\beta(\mathbf{s}^n)p(\mathbf{s}^n|\beta)p(\mathbf{c}^n|\mathbf{s}^n,\beta)} = \\ = \frac{Pr_\phi(\mathbf{s}^n)p(\mathbf{x}^n|\phi)}{Pr_\phi(\mathbf{s}^n)p(\mathbf{x}^n|\phi) + Pr_\beta(\mathbf{s}^n)p(\mathbf{x}^n|\beta)} \tag{5.19}$$

Por lo tanto, para aplicar este clasificador, además de los modelos estimados para el fondo y el primer plano sólo necesario el cálculo de $Pr_\phi(\mathbf{s}^n)$.

En las zonas de la imagen en las que no se sitúe ninguna partícula predicha no se dispone de información a priori relativa a la velocidad o a la trayectoria de los objetos móviles. En estas regiones, al no disponerse de ninguna información a priori, se debe establecer que $Pr_\phi(\mathbf{s}^n) = Pr_\beta(\mathbf{s}^n) = \frac{1}{2}$. Sin embargo, en las zonas en las que los píxeles estén cubiertos por dichas partículas predichas, el valor de $Pr_\phi(\mathbf{s}^n)$ debe ser superior a $\frac{1}{2}$ (con menor o mayor valor en función de la cantidad de partículas situadas sobre los píxeles y de la distancia entre los centros de las partículas y dichos píxeles). Atendiendo a estos criterios, se ha definido $Pr_\phi(\mathbf{s}^n)$ como:

$$Pr_\phi(\mathbf{s}^n) = \begin{cases} \frac{1}{2}, & N_p \leq N_n \\ \frac{1}{N_p}\sum_{i=1}^{N_p} G_i(\mathbf{s}^n), & N_p > N_n \end{cases} \tag{5.20}$$

donde N_p es el número de partículas predichas que afectan a la posición espacial analizada, $G_i(\mathbf{s}^n)$ es el valor del perfil de la partícula i-ésima sobre dicha posición, y N_n es un umbral de ruido que evita tener en cuenta las medidas ruidosas.

Tal y como se describe en el apéndice C del presente documento, se han utilizado elipses para representar espacialmente las partículas predichas y se les ha asignado un perfil gaussiano que se define como:

$$G_i(\mathbf{s}^n) = \exp\left(-\frac{1}{2}\left(\frac{(\mathbf{s}^n(1) - h_i^n)^2}{(\eta a_i^n)^2} + \frac{(\mathbf{s}^n(2) - w_i^n)^2}{(\eta b_i^n)^2}\right)\right) \tag{5.21}$$

donde (h_i^n, w_i^n) son las coordenadas en las que se sitúa el centro de la elipse i-ésima, (a_i^n, b_i^n) son los ejes de dicha elipse, y η un factor que permite garantizar que los valores de las gaussianas en el contorno de las elipses es igual a $\frac{1}{2}$. Por lo tanto, el valor de dicho factor es $\eta \simeq 0,85$.

De esta forma, todos los píxeles que estén cubiertos por más de N_n partículas tendrán un valor de $Pr_\phi(\mathbf{s}^n)$ que estará comprendido entre $\frac{1}{2}$ y 1 (mayor valor para menor distancia espacial a los centros de las elipses que los cubren). Por otro lado, el valor de $Pr_\phi(\mathbf{s}^n)$ para el resto de los píxeles de la imagen (aquellos de los que no se dispone de información que permita decantarse por uno u otro valor de probabilidad a priori) tendrán asignado el valor $Pr_\phi(\mathbf{s}^n) = \frac{1}{2}$.

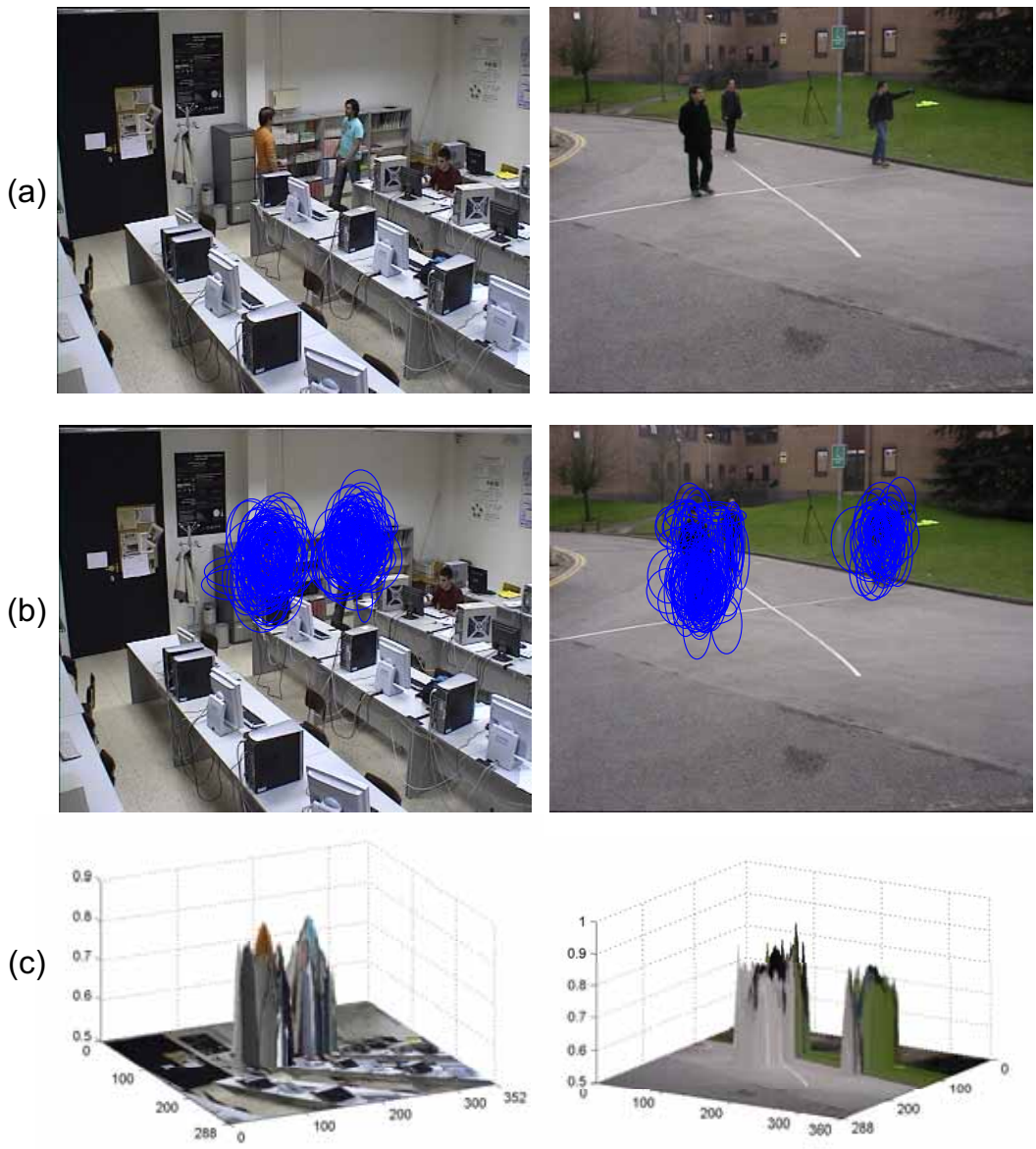


Figura 5.9: Probabilidades a priori obtenidas de las predicciones del filtro de partículas. (a) Imágenes originales. (b) Partículas predichas sobre las imágenes originales. (c) Representación tridimensional de las probabilidades obtenidas de las partículas predichas, $Pr_{\phi}(\mathbf{s}^n)$, sobre las imágenes originales (nótese que el menor valor de estas probabilidades es $0,5$).

La figura 5.9 presenta dos ejemplos que permiten apreciar la influencia de las partículas predichas en el cálculo de las probabilidades a priori. La primera fila de imágenes (figura 5.9.a) muestra dos imágenes originales, pertenecientes a dos secuencias con distinto número de objetos móviles y en escenarios muy diferentes. En la segunda fila de imá-

genes (figura 5.9.b) se han representado los conjuntos de partículas predichas obtenidas tras la aplicación del filtro de partículas sobre las detecciones obtenidas para las imágenes originales. Por último, la tercera y última fila de imágenes (figura 5.9.c) muestra las representaciones tridimensionales de los valores de $Pr_{\phi}(\mathbf{s}^n)$ sobre las imágenes originales. Dichas representaciones permiten apreciar que los valores de $Pr_{\phi}(\mathbf{s}^n)$ son superiores a $\frac{1}{2}$ únicamente en las zonas de las imágenes cubiertas por más de N_n partículas.

5.6. Estimación dinámica del ancho de los *kernels*

Recordando lo dicho en la sección 5.3.1, para la correcta estimación de las funciones de densidad, la elección de matrices de escala que determinen un ancho adecuado para los *kernels* utilizados en la estimación es de gran importancia.

Para obtener resultados de gran calidad las estrategias más complejas utilizan matrices completamente parametrizadas (Mittal y Paragios, 2004) (Wan y Wang, 2008) (Tavakkoli et al., 2009). Sin embargo, para construir estas matrices requieren tener almacenada una gran cantidad de datos y realizar un elevado número de operaciones por píxel, lo cual resulta en una elevada carga computacional y de memoria, disminuyendo muy notablemente su eficiencia.

Otras estrategias (Elgammal et al., 2002) (Martel-Brisson y Zaccarin, 2008), buscando un compromiso razonable entre calidad y eficiencia, asumen independencia entre todas las componentes utilizadas para representar la información de los píxeles y, de ese modo, pueden hacer uso de matrices de escala diagonales:

$$\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2 \dots \sigma_{D+2}^2) \tag{5.22}$$

en las que σ_i^2 determina el ancho correspondiente a la componente i-ésima de los *kernels*.

Por otro lado están las estrategias que, de forma similar a la aquí presentada, combinan un modelado del fondo con un modelo del primer plano haciendo uso de la información espacial de las muestras (Sheikh y Shah, 2005) (Zhang y Yang, 2008). Estas estrategias, para mantener sus requisitos computacionales y debido a la falta de métodos capaces de estimar dinámicamente las matrices de escala de los *kernels* que tienen en cuenta la información espacial de los píxeles, utilizan matrices diagonales con valores fijos. De esta forma se ahorran el coste computacional asociado a la estimación de dichas matrices, consiguiendo reducir el tiempo de procesamiento. Sin embargo, estas estrategias, para obtener resultados con la calidad suficiente deben ajustar manualmente los parámetros de estas matrices en función de las características de cada secuencia. Además, dado que las características de una secuencia pueden cambiar a lo largo de la misma, utilizando las mismas matrices a lo largo de toda la secuencia no es posible mantener la calidad de las detecciones obtenidas.

En la figura 5.10 se han representado algunos resultados, obtenidos mediante la aplicación de distintas estrategias que hacen uso de diferentes métodos de estimación de los anchos de los *kernels*. Estos resultados se han obtenido para una imagen perteneciente a una secuencia en la que aparece un único objeto móvil (figura 5.10.a). Los resultados mostrados en la segunda fila de imágenes de la figura se han obtenido utilizando matrices diagonales

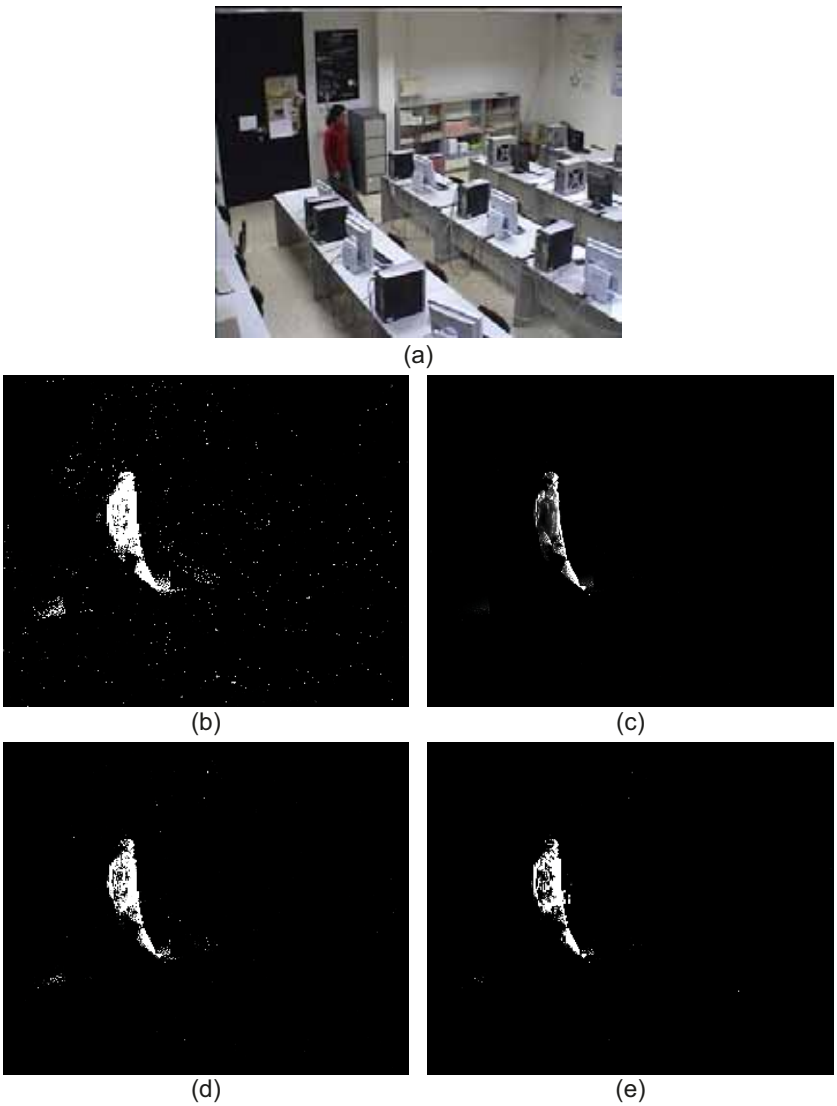


Figura 5.10: Detecciones obtenidas con distintas matrices de escala. (a) Imagen original. (b) Con valores fijos y pequeños. (c) Con valores fijos y altos. (d) Con matrices diagonales (Elgammal et al., 2002). (e) Con matrices completamente parametrizadas (Tavakkoli et al., 2009).

con parámetros fijos: de menor valor en el caso de la figura 5.10.b; y de mayor valor en el caso de la figura 5.10.c. Como resultado de la utilización de los *kernels* más estrechos (figura 5.10.b) se ha obtenido una detección en la que, aunque el objeto móvil ha sido correctamente clasificado, la calidad de falsas detecciones es elevada. Por otro lado, en el resultado correspondiente a la utilización de *kernels* más anchos (figura 5.10.c), el número de falsas detecciones es mucho menor. Sin embargo, en este caso el objeto móvil no ha sido detectado correctamente. Los resultados mostrados en la tercera fila de imágenes de la figura se han

obtenido utilizando estrategias de modelado que estiman dinámicamente las matrices de escala. El mostrado en la figura 5.10.d se ha logrado utilizando matrices diagonales (Elgammal et al., 2002), mientras que el mostrado en la figura 5.10.e se ha obtenido con la aplicación de matrices parametrizadas por completo (Tavakkoli et al., 2009). En ambos casos las detecciones obtenidas muestran un compromiso aceptable entre el número de falsas detecciones y el de píxeles móviles no detectados, siendo algo mejor el resultado obtenido con el método propuesto en (Tavakkoli et al., 2009).

En esta sección se describen dos estrategias para la estimación dinámica de las matrices de escala utilizadas en los métodos de modelado no paramétrico descritos en la sección 5.4, en los cuales se utiliza tanto información temporal como espacial. Para estimar las matrices de escala utilizadas en el modelado del fondo se ha desarrollado una eficiente y creativa estrategia basada en el análisis estadístico de las muestras de referencia, pesadas de acuerdo a su localización espacial dentro de las imágenes. En el caso de las matrices de escala utilizadas en el modelado del primer plano se elaborado una innovadora estrategia, basada en la agrupación de muestras de referencia con *Mean-Shift* (Cheng, 1995), que permite estimar el ancho de cada uno de los modos a los que pertenecen dichas muestras. A continuación se describen estas dos estrategias de estimación.

5.6.1. Estimación dinámica del ancho de las matrices de escala utilizadas el modelado del fondo

Para llevar a cabo el modelado del fondo se ha decidido determinar el ancho de los *kernels* mediante matrices diagonales. De este modo se consigue mejorar notablemente la calidad de los resultados con respecto al caso de utilizar matrices con valores fijos, sin que el coste computacional y de memoria suponga un aumento tan elevado como el que resultaría de la utilización de matrices parametrizadas por completo. Dichas matrices se definen como:

$$\Sigma_\beta = \text{diag}(\sigma^2_{\beta_1}, \sigma^2_{\beta_2} \ldots \sigma^2_{\beta_D}, \sigma^2_{\beta_H}, \sigma^2_{\beta_W}) \tag{5.23}$$

en las que las D primeras componentes son las componentes que determinan el ancho correspondiente a las D características de apariencia de los píxeles, y las dos últimas determinan el ancho correspondiente a las características espaciales: $\sigma^2_{\beta_H}$ para las filas y $\sigma^2_{\beta_W}$ para las columnas. Haciendo uso de matrices de este tipo, la ecuación 5.7 puede volver a escribirse como:

$$p(\mathbf{x}^n|\beta) = \frac{1}{N_\beta (2\pi)^{\frac{D+2}{2}}} \sum_{i=1}^{N_\beta} \prod_{j=1}^{D+2} \frac{1}{(\Sigma_\beta(j,j))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^n(j) - \mathbf{x}^i_\beta(j))^2}{\Sigma_\beta(j,j)}\right) \tag{5.24}$$

Dado que la información proporcionada por el conjunto de muestras de referencia, $\{\mathbf{x}^i_\beta\}_{i=1}^{N_\beta}$, está uniformemente distribuida en el espacio, se ha decidido que el ancho de los *kernels* en sus componentes espaciales tenga un valor fijo y de idéntico valor tanto en las filas como en las columnas, $\sigma_{\beta_H} = \sigma_{\beta_H} = \sigma_{\beta_s}$. Cuanto mayor sea el valor de σ_{β_s}, mejor será el modelado de las zonas ruidosas del fondo. Sin embargo, valores demasiado altos pueden suponer un

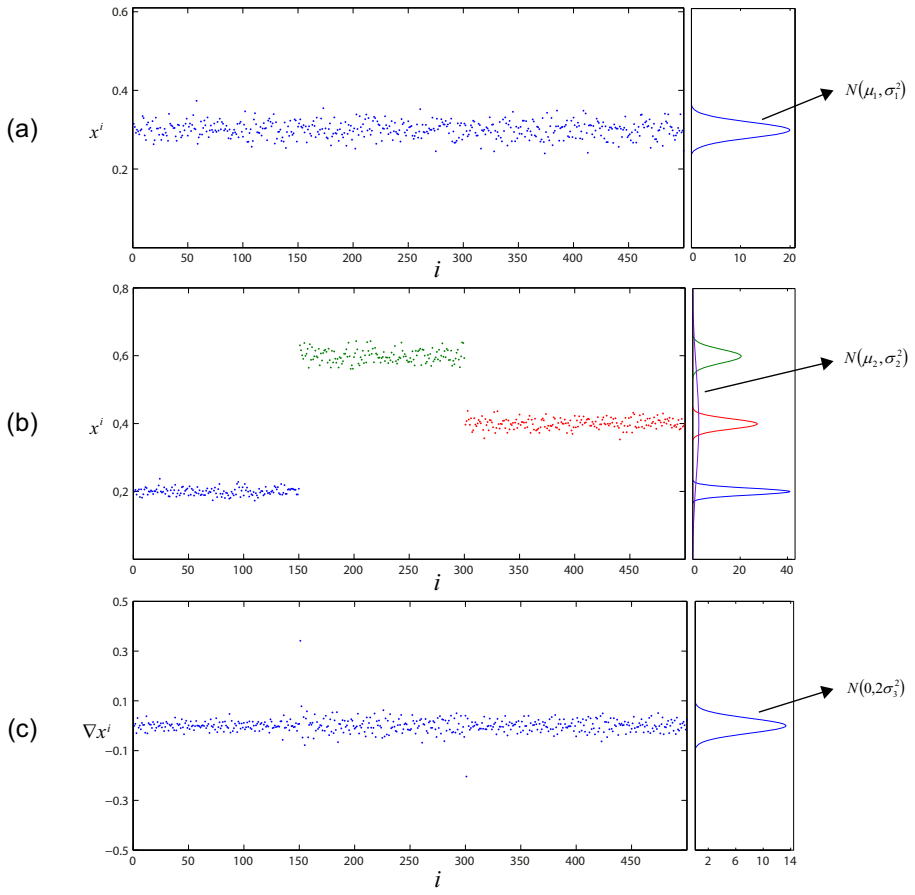


Figura 5.11: (a) A la izquierda, conjunto de muestras pertenecientes a una distribución unimodal; a la derecha, *kernel* gaussiano extraído de dicha distribución. (b) A la izquierda, conjunto de muestras pertenecientes a una distribución multimodal; a la derecha, gaussianas asociadas a cada uno de los modos (en azul, verde y rojo) y *kernel* gaussiano extraído de la distribución completa (en morado). (c) A la izquierda, conjunto de diferencias entre las muestras consecutivas de la distribución representada en (b); a la derecha, gaussiana asociada a dichas diferencias.

coste computacional demasiado elevado y, además, dan lugar a una peor clasificación de los objetos móviles. En la sección de resultados se analiza detalladamente la calidad de las detecciones obtenidas en función del valor asignado a este ancho espacial.

Por otro lado, puesto que la información correspondiente a las D características de apariencia no está uniformemente distribuida, es necesario estimar los anchos más adecuados para modelar sus variaciones. Para obtener estas estimaciones se ha utilizado una estrategia que parte de la idea propuesta en (Elgammal et al., 2002), la cual estima las matrices de escala a partir de la mediana, calculada para cada componente por separado, de los valores absolutos de las diferencias entre muestras consecutivas situadas en la misma posición

espacial.

Considérese un conjunto de muestras de referencia asociadas al píxel p^n, extraídas de los píxeles situados en la misma posición espacial que p^n en imágenes anteriores. En las situaciones más sencillas, en las que dicho píxel no haya sufrido variaciones significativas, todas las muestras formarán parte de una distribución constituida por un único modo que puede aproximarse mediante una gaussiana definida por la media y la varianza de dichas muestras de referencia. En estos casos, utilizando *kernels* gaussianos con un ancho dado por la varianza de las muestras de referencia, la distribución a la que pertenecen dichas muestras será adecuadamente estimada. En la figura 5.11.a se ha representado un ejemplo de esta situación, para el caso de un conjunto de 500 muestras de referencia unidimensionales. En la imagen de la izquierda se han representado los valores de las muestras de referencia, mientras que en la imagen de la derecha se ha dibujado una gaussiana, $N(\mu_1, \sigma_1^2)$, caracterizada por la media y la varianza del conjunto de muestras de referencia. En situaciones como las de este ejemplo, si en el modelado no paramétrico del fondo se utilizan matrices de escala con un ancho definido por la varianza de las muestras de refencia, $\Sigma_\beta = \sigma_1^2$, será posible obtener una buena estimación de la distribución de probabilidad a la que pertenecen las muestras.

Sin embargo, si las muestras de referencia se corresponden con un píxel que ha sufrido cambios o que ha sido ocluido por otro objeto, no formarán parte de una distribución con un único modo, sino que pertenecerán a una distribución multimodal. En la figura 5.11.b se muestra un ejemplo de esta situación, para el caso de 500 muestras de referencia pertenecientes a una distribución constituida por 3 modos. En la imagen de la izquierda se han representado los valores las muestras, asignándoles distinto color en función del modo al que pertenecen. En la imagen de la derecha se han representado las tres gaussianas asociadas a cada uno de los modos de la distribución (con colores similares a los de las muestras con las que se corresponden), caracterizadas por la media y la varianza de cada uno de los 3 modos. Además de estas 3 gaussianas se ha representado una cuarta gaussiana (en color morado), $N(\mu_2, \sigma_2^2)$, en la que μ_2 y σ_2^2 son la media y la varianza de las 500 muestras de referencia. En este caso, si al igual que en el ejemplo anterior, el modelado no paramétrico se lleva a cabo mediante *kernels* gaussianos con matrices de escala definidas por la varianza del conjunto de todas las muestras de referencia, $\Sigma_\beta = \sigma_2^2$, la estimación obtenida será de baja calidad ya que, al utilizarse *kernels* demasiado anchos, no será posible diferenciar los 3 modos de los que consta la distribución real.

Para modelar correctamente este tipo de distribuciones multimodales se han de utilizar gaussianas con un ancho que se adapte lo mejor posible a cada uno de los modos por separado. Para estimar este ancho, el método presentado en (Elgammal et al., 2002) parte de la hipótesis de que cada uno de los modos de la distribución que se desea estimar se adapta a una misma gaussiana, $N(\mu_3, \sigma_3^2)$, y construye una distribución de diferencias entre muestras consecutivas, $N(0, 2\sigma_3^2)$, que permite comparar los datos de todos los modos de la distribución original y obtener una varianza que sea representativa del conjunto total de las muestras de referencia. En el ejemplo mostrado en la figura 5.11.c se ha representado la distribución de diferencias que se corresponde con la distribución multimodal representada en la figura 5.11.b. En la imagen de la izquierda se pueden ver los valores de las diferencias entre las muestras consecutivas de la distribución original y en la de la derecha se ha

representado la gaussiana definida por la media de dicha distribución de diferencias y una desviación típica obtenida como:

$$\sigma_3 = \frac{m}{0,68\sqrt{2}} \tag{5.25}$$

donde m es la mediana del valor absoluto de las diferencias. Dado que las distribuciones de diferencias son simétricas esta mediana equivale a su percentil 75 y, por lo tanto, debe ser tal que verifique que:

$$Pr(N(0, 2\sigma_3^2) > m) = 0{,}25 \tag{5.26}$$

Estimando la desviación típica a partir de la mediana, al ser esta un estimador robusto que evita la influencia negativa de los datos atípicos de una distribución, se consigue no tener en cuenta las diferencias entre las muestras pertenecientes a distintos modos (por ser estas mucho menos frecuentes que las diferencias entre muestras pertenecientes a un mismo modo), obteniéndose así estimaciones muy precisas que permiten asignar a los *kernels* anchos muy adecuados.

A diferencia que en los ejemplos mostrados para explicar el funcionamiento de este método de estimación, en los que todas las muestras de referencia están situadas en la misma posición espacial, la estrategia de modelado propuesta en este capítulo utiliza un conjunto de muestras de referencia extraídas de píxeles situados en distintas coordenadas. Por lo tanto, es necesario adaptar el método previamente descrito, de forma que tenga en cuenta la localización espacial de los píxeles.

Sea el conjunto de muestras de referencia $\{\mathbf{x}_\beta^i\}_{i=1}^{N_\beta}$, asociadas al píxel p^n. Dicho conjunto puede ser dividido en subconjuntos, de forma que las muestras de cada subconjunto tengan idénticas componentes espaciales (se correspondan con píxeles situados en las mismas coordenadas). Siguiendo el razonamiento expuesto anteriormente para estimar la varianza de un conjunto de muestras pertenecientes a una distribución multimodal, de cada uno de estos subconjuntos se puede obtener una distribución de diferencias que permita estimar una matriz de escala adecuada para el análisis de dichos subconjuntos. Sin embargo, para obtener una matriz de escala que sea representativa del conjunto total de muestras de referencia, se deben considerar de forma conjunta todas las distribuciones de diferencias obtenidas de estos subconjuntos: la mediana utilizada en la ecuación 5.25 debe ser calculada teniendo en cuenta las distribuciones de diferencias obtenidas de todos los subconjuntos. Por otro lado, también debe tenerse en cuenta que la influencia de cada muestra de referencia en la estimación del modelo del fondo decrece a medida que aumenta su distancia espacial con p^n. Por lo tanto, la matriz de escala deberá adaptarse mejor a las distribuciones de diferencias con menor distancia espacial a p^n. Para conseguirlo, se ha decidido asignar un peso a cada una de estas distribuciones de diferencias en función de la distancia entre sus muestras y p^n. Dichos pesos se han obtenido mediante la aplicación de gaussianas espaciales similares a las utilizadas en el modelado no paramétrico del fondo y se obtienen como:

$$w_\sigma = \lambda \ \exp\left(-\frac{(\Delta h)^2 + (\Delta w)^2}{2\sigma_s^2}\right) \tag{5.27}$$

donde λ es un factor de normalización para los factores de peso y $(\Delta h, \Delta w)$ son las distancias, en filas y columnas, entre cada subconjunto de muestras y el píxel p^n.

5.6.2. Estimación dinámica del ancho de las matrices de escala utilizadas el modelado del primer plano

Para modelar el primer plano, al igual que en el caso del modelado del fondo y buscando un compromiso entre eficiencia y calidad, se han utilizado *kernels* gaussianos caracterizados por matrices de escala diagonales:

$$\Sigma_\phi = \text{diag}(\sigma^2_{\phi_1}, \sigma^2_{\phi_2} \ldots \sigma^2_{\phi_D}, \sigma^2_{\phi_H}, \sigma^2_{\phi_W}) \tag{5.28}$$

en las que las D primeras componentes determinan el ancho correspondiente a las características de apariencia de las muestras de referencia y las dos últimas determinan el de sus características espaciales: $\sigma^2_{\phi_H}$ para las filas y $\sigma^2_{\phi_W}$ para las columnas. De este modo, la ecuación 5.9 puede volver a escribirse como:

$$p(\mathbf{x}^n|\phi) = \alpha\gamma + \frac{(1-\alpha)}{N_\phi (2\pi)^{\frac{D+2}{2}}} \sum_{i=1}^{N_\phi} \prod_{j=1}^{D+2} \frac{1}{(\Sigma_\phi(j,j))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^n(j) - \mathbf{x}^i_\phi(j))^2}{\Sigma_\phi(j,j)}\right) \tag{5.29}$$

Para obtener una estimación precisa de la función densidad de probabilidad del primer plano se debe hacer uso de *kernels* que se adapten lo mejor posible a todos los modos de dicha distribución, sin ser demasiado anchos ni demasiado estrechos. Por lo tanto, si a cada una de las muestras de referencia utilizadas en el modelado se le asigna una matriz de escala que se adapte al modo al que pertenece dicha muestra, será posible estimar un modelo preciso del primer plano.

Para alcanzar este propósito se ha planteado una estrategia que, haciendo uso de *Mean-Shift* (Comaniciu y Meer, 2002), permite agrupar en regiones homogéneas las N_ϕ muestras de referencia utilizadas en el modelado del primer plano y asignar a cada muestra una matriz de escala definida por la varianza de la región homogénea a la que pertenezca.

Mean-Shift es un algoritmo iterativo y no paramétrico que, a partir de un conjunto de muestras multidimensionales, permite localizar los máximos locales de la función densidad de probabilidad a la que pertenecen dichas muestras (Nieto et al., 2011). En cada iteración, *Mean-Shift* parte de una muestra y, dentro de un espacio de búsqueda definido en el entorno de dicha muestra, se desplaza hacia la región de muestras más densa. A continuación se describe el modo de funcionamiento de este algoritmo.

Sea el conjunto de muestras asociadas a los píxeles de las regiones de primer plano, $\{\mathbf{x}^i_\phi\}_{i=1}^{N_\phi}$, obtenidas a los largo de las N_{ϕ_N} imágenes previas a la actual y que, gracias a la estrategia de seguimiento descrita en la sección 5.5, han sido actualizadas de imagen a imagen. La función densidad de probabilidad a la que pertenecen estas muestras, $f_\phi(\mathbf{x})$, se

puede estimar de forma no paramétrica, tal como se ha descrito en la sección 5.3, como:

$$\hat{f}(\mathbf{x}) = \frac{1}{N_\phi} \sum_{i=1}^{N_\phi} \frac{1}{|\Sigma_\phi|^{\frac{1}{2}}} K_{MS}\left(\frac{\mathbf{x} - \mathbf{x}_\phi^i}{\Sigma_\phi^{\frac{1}{2}}}\right) \tag{5.30}$$

Dado que se está considerando que las $(D+2)$ componentes utilizadas son independientes entre sí, el *kernel* que aparece en esta expresión, K_{MS}, puede ser expresado como un producto de *kernels* simétricos de una dimensión, $K_1(\mathbf{x}(j))$ (Comaniciu y Meer, 2002):

$$K_{MS}(\mathbf{x}) = \prod_{j=1}^{D+2} K_1(\mathbf{x}(j)) \tag{5.31}$$

Para simplificar la notación (Comaniciu et al., 2001), en vez de utilizarse directamente estos *kernels* se utiliza su perfil, $k : [1, \infty) \rightarrow \mathbb{R}$, el cual verifica que $K_{MS}(\mathbf{x}) = C_{MS}\ k(\|\mathbf{x}\|^2)$, donde C_{MS} es una constante de normalización.

Los modos de $f_\phi(\mathbf{x})$ estarán situados en los ceros del gradiente $\nabla f_\phi(\mathbf{x})$, el cual puede ser estimado como:

$$\nabla \hat{f}_\phi(\mathbf{x}) = A_{MS}\left[\sum_{i=1}^{N_\phi} G_{MS}(\mathbf{x} - \mathbf{x}^i)\right]\left[\left(\sum_{i=1}^{N_\phi} G_{MS}(\mathbf{x} - \mathbf{x}^i)\right)^{-1}\left(\sum_{i=1}^{N_\phi} G_{MS}(\mathbf{x} - \mathbf{x}^i)\mathbf{x}^i\right) - \mathbf{x}\right] \tag{5.32}$$

donde A_{MS} es una matriz de $(D+2)\times(D+2)$ componentes, constituida por varias constantes y por la matriz de escala, Σ_ϕ, que determina el ancho de los *kernels*:

$$A_{MS} = \frac{2C_{MS}}{N_\phi |\Sigma_\phi|^{\frac{1}{2}}} \Sigma_\phi^{-2} \tag{5.33}$$

y $G_{MS}(\mathbf{x})$ es una matriz de $(D+2) \times (D+2)$ componentes definida como:

$$G_{MS}(\mathbf{x}) = \text{diag}\left[-k'\left(\frac{\mathbf{x}^2(1)}{\Sigma_\phi^2(1,1)}\right), -k'\left(\frac{\mathbf{x}^2(2)}{\Sigma_\phi^2(2,2)}\right) \ldots, -k'\left(\frac{\mathbf{x}^2(D+2)}{\Sigma_\phi^2(D+2, D+2)}\right)\right] \tag{5.34}$$

siendo k' la derivada del perfil de los *kernels* unidimensionales.

El último factor definido en la ecuación 5.32 es el vector que representa la diferencia entre la media de los datos que se evalúan y el centro del *kernel*:

$$m_{MS}(\mathbf{x}) = \frac{\sum_{i=1}^{N_\phi} G_{MS}(\mathbf{x} - \mathbf{x}^i)\mathbf{x}^i}{\sum_{i=1}^{N_\phi} G_{MS}(\mathbf{x} - \mathbf{x}^i)} - \mathbf{x} \tag{5.35}$$

Partiendo de cada una de las muestras del conjunto $\{\mathbf{x}_\phi^i\}_{i=1}^{N_\phi}$, aplicando este vector de forma iterativa, es posible localizar el modo asociado a cada una de ellas y, por lo tanto, agruparlas en función del modo al que pertenecen.

Para llevar a cabo esta agrupación basada en *Mean-Shift* se ha hecho uso de *kernels* de *Epanechnikov*, los cuales se definen como:

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2c_D}(D+2)(1-|\mathbf{x}|^2) & \text{si } |\mathbf{x}| \leq 1 \\ 0 & \text{si } |\mathbf{x}| > 1 \end{cases} \tag{5.36}$$

y cuyo perfil es:

$$k_E(\mathbf{x}) = \begin{cases} \frac{1}{2c_D}\frac{D+2}{C_{MS}}(1-\mathbf{x}) & \text{si } |\mathbf{x}| \leq 1 \\ 0 & \text{si } |\mathbf{x}| > 1 \end{cases} \tag{5.37}$$

donde c_D es el volumen de una esfera D-dimensional unitaria. Mientras que la derivada del perfil de un *kernel* gaussiano es una función gaussiana, la derivada del perfil del *kernel* de *Epanechnikov* es una función uniforme y, por lo tanto, permite obtener muy eficientemente el vector definido en la expresión 5.35, calculando la diferencia entre la muestra actual, $\mathbf{x}$, y el valor medio del subconjunto de muestras, Ψ, consideradas por el *kernel* en cada iteración. Por lo tanto, utilizando dicho *kernel*, el desplazamiento definido en la ecuación 5.35 puede ser calculado de forma muy eficiente como:

$$m_{MS}(\mathbf{x}) = \frac{1}{|\Psi|}\sum_{i\in\Psi}(\mathbf{x}_\phi^i - \mathbf{x}), \text{ donde } \Psi = \left\{ i \in (1,\ldots,N_\phi) \;/\; \left| \frac{\mathbf{x}-\mathbf{x}_\phi^i}{\Sigma_\phi^{\frac{1}{2}}} \right| \leq 1 \right\} \tag{5.38}$$

donde $|\Psi|$ es el cardinal del conjunto Ψ.

Una vez agrupados los píxeles, a cada uno se le asigna una matriz como la definida en la ecuación 5.28. Los elementos de estas matrices vendrán dados por la varianza del conjunto de muestras de la región homogénea a la que pertenezca cada píxel. Estas matrices, además de utilizarse para definir el ancho de los *kernels* utilizados en el modelado no paramétrico del primer plano de la imagen actual, también serán utilizadas para determinar el ancho de los *kernels* utilizados en el proceso de agrupación con *Mean-Shift* efectuado sobre la siguiente imagen.

En el caso de que parte del conjunto de muestras del primer plano contenga datos nuevos, las matrices de escala asociadas a dichos datos, Σ_ϕ, se construirán a partir de las varianzas del conjunto de muestras de la región móvil a la que pertenezcan.

Se ha de tener en cuenta que, al contrario que en el modelado del fondo, en el que se utiliza un *kernel* con distinto ancho para cada píxel de la imagen actual (*balloon estimator*), en el modelado del primer plano se está utilizando distinto ancho en cada muestra de referencia y, por lo tanto, se está haciendo uso de un *sample-point estimator*. Además, también se debe tener en consideración que la estrategia propuesta, al contrario que la mayor parte de las estrategias que hacen uso de *Mean-Shift*, utiliza *kernels* con distinto ancho en cada dimensión (Han et al., 2007), permitiendo la obtención de resultados más precisos.

Por último se ha de mencionar que, debido a que *Mean-Shift* se aplica sobre un conjunto reducido de muestras (sólo aquellas clasificadas como parte del primer plano a lo largo de

las N_{ϕ_N} imágenes anteriores) y gracias a que se están utilizando *kernels* que permiten trabajar muy eficientemente (*kernels* de *Epanechnikov*), el coste computacional asociado a la estimación de estas matrices de escala es asumible. Además, dado que únicamente se debe definir una matriz de escala para cada región homogénea resultante del proceso de agrupación, la cantidad de memoria requerida por la estrategia resulta despreciable. El porcentaje de carga computacional que supone esta etapa de estimación se analiza más detalladamente en el capítulo 6.

5.7. Resultados

Para evaluar la calidad del sistema de detección presentado en este capítulo se han analizado los resultados obtenidos tras su aplicación sobre las secuencias de la base de datos descrita en el apéndice A.

En primer lugar, en la sección 5.7.1, se muestran las mejoras obtenidas mediante la estrategia propuesta para modelar el fondo. A continuación, en la sección 5.7.2, se analizan las mejoras logradas con la estrategia propuesta para el modelado del primer plano. Por último, en la sección 5.7.3, se presentan los resultados correspondientes a las detecciones finales logradas, obtenidas a partir de ambos modelados y de la probabilidad a priori estimada.

Para llevar a cabo la evaluación cuantitativa de los resultados presentados en esta sección se han utilizado los tres porcentajes (*Recall*, *Precision* y F) descritos en la sección 3.7, calculados del modo descrito en la sección 4.4. Como información de apariencia de los píxeles se han considerado únicamente sus componentes de color RGB, por lo que en las ecuaciones descritas a lo largo del capítulo se ha utilizado $D = 3$.

Las estrategias de modelado propuestas se han comparado, en términos de calidad y eficiencia computacional, con las propuestas realizadas en dos trabajos que proponen distintas alternativas para llevar a cabo el modelado no paramétrico del fondo y del primer plano. La primera de estas estrategias (Sheikh y Shah, 2005) es una importante referencia dentro del campo de la detección de objetos móviles mediante técnicas de modelado no paramétrico. Esta estrategia, al igual que la descrita en el presente capítulo, también hace uso de información espacial para llevar a cabo los modelados del fondo y del primer plano, logrando resultados de gran calidad. Sin embargo, debido a que no actualiza las posiciones de los objetos móviles previamente detectados, no proporciona resultados suficientemente buenos y el coste computacional que requiere es muy elevado. Además, para evitar un aumento todavía mayor del coste computacional, utiliza *kernels* con un ancho fijo, lo cual reduce su capacidad para obtener resultados satisfactorios en secuencias con distintas características. La segunda estrategia (Zhang y Yang, 2008), basada en la anterior, propone la actualización de las posiciones de los objetos móviles a partir de un análisis basado en histogramas. Gracias a esta actualización es capaz de mejorar los resultados de (Sheikh y Shah, 2005). Sin embargo, esta estrategia no es aplicable a secuencias con múltiples objetos móviles y, además, hace uso de una etapa de pre-detección que reduce notablemente su eficiencia computacional.

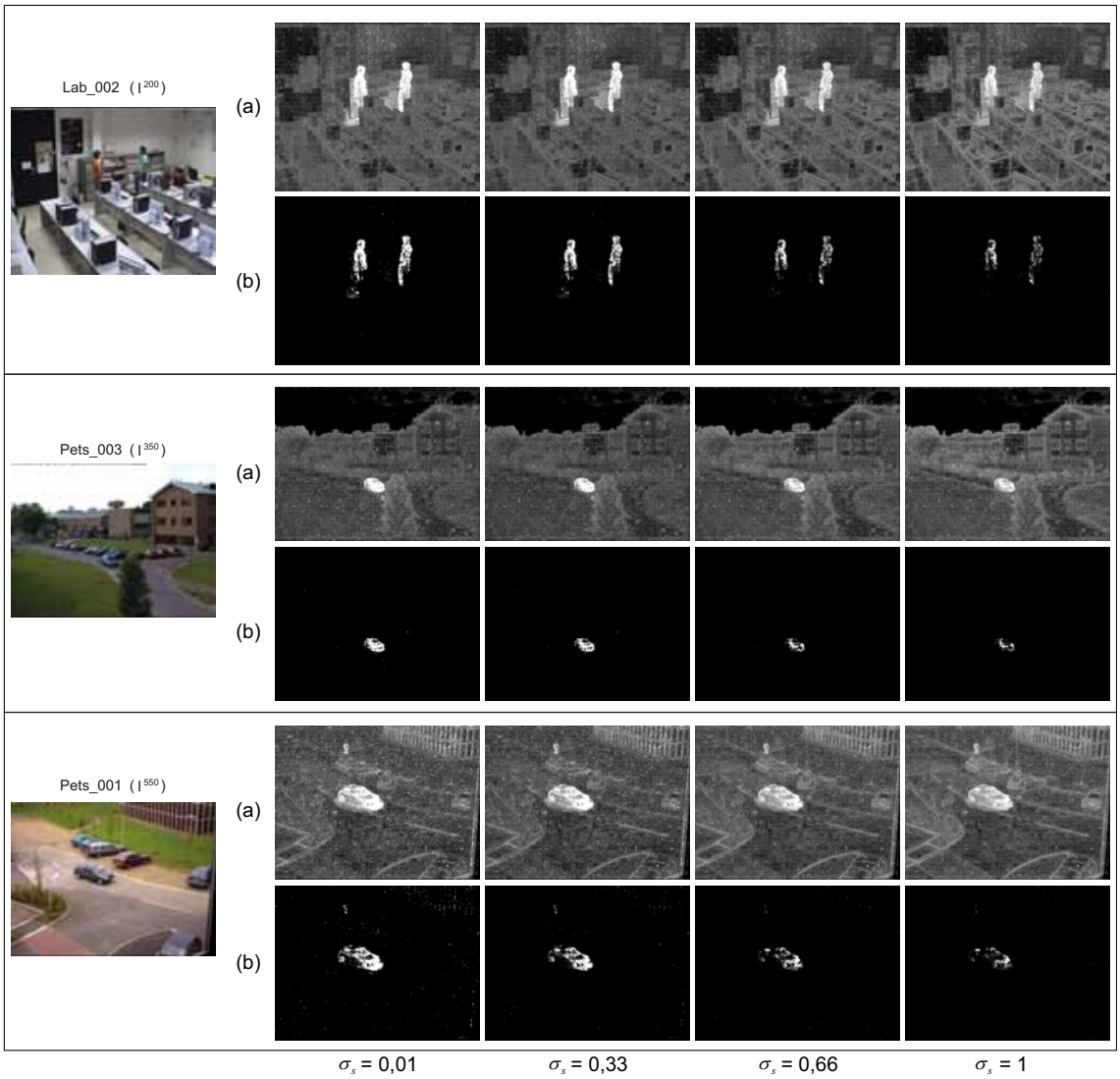


Figura 5.12: Análisis de los resultados en función del ancho espacial, σ_s, utilizado en el modelado del fondo. (a) Logaritmo-negativo de la función densidad probabilidad del fondo. (b) Detecciones finales.

5.7.1. Modelado del fondo

En esta sección se analiza la calidad de la estrategia propuesta para llevar a cabo el modelado del fondo, el cual se ha obtenido haciendo uso de $N_{\beta_N} = 150$ imágenes de referencia. Las detecciones finales obtenidas a partir de dicha estrategia y mostradas en esta sección se han obtenido aplicando el clasificador bayesiano descrito en la ecuación 5.11, en el que la verosimilitud del primer plano, $p(\mathbf{x}^n|\phi)$, se ha considerado uniforme y a las probabilidades a priori de ambas clases se les ha asignado el mismo valor, $Pr(\beta) = Pr(\phi) = \frac{1}{2}$.

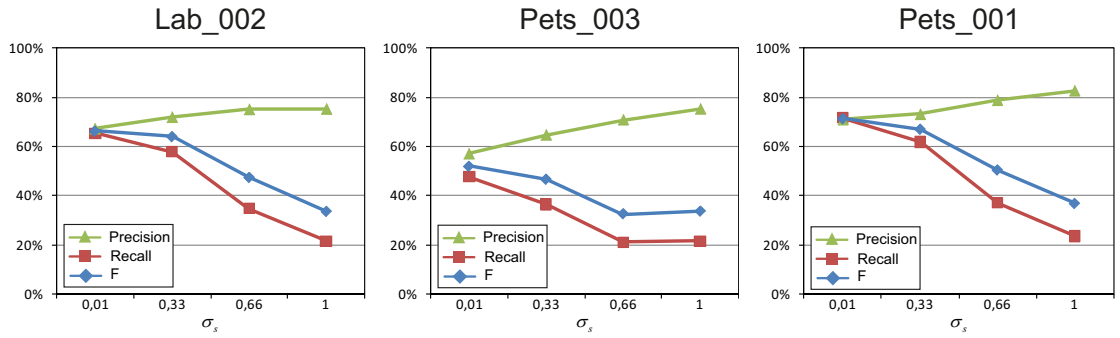


Figura 5.13: Resultados cuantitativos en función del ancho espacial de los *kernels* utilizados en el modelado del fondo.

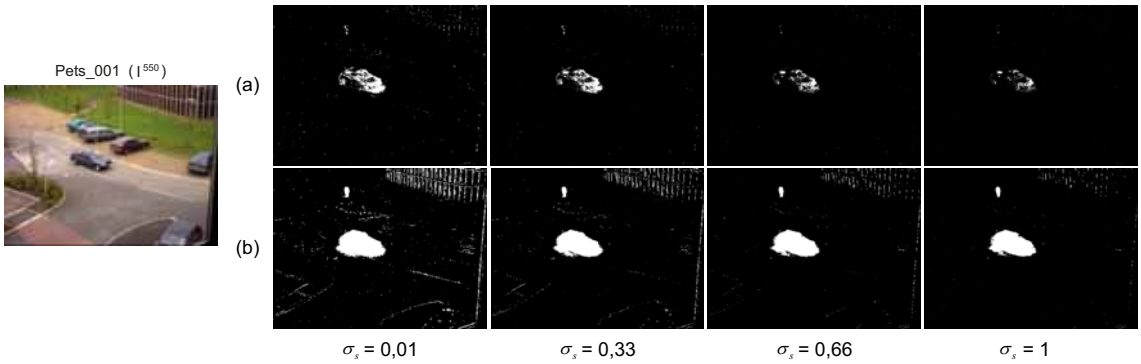


Figura 5.14: Resultados obtenidos a partir de distintos datos de referencia en el modelado del fondo. (a) Utilizando toda la información de las imágenes anteriores. (b) Utilizando únicamente la información de los píxeles etiquetados como parte del fondo.

En primer lugar se ha analizado la calidad de los resultados en función del valor asignado al parámetro σ_s, el cual determina el ancho espacial de los *kernels* gaussianos utilizados en el modelado del fondo. La figura 5.12 muestra algunos de los resultados obtenidos sobre tres secuencias de distintas características, utilizando distintos valores de σ_s. Los valores de *Recall*, *Precision* y F que se corresponden con estos resultados aparecen representados en las gráficas de la figura 5.13.

Analizando los resultados mostrados en estas dos figuras es posible apreciar que cuanto mayor es el ancho de los *kernels* en sus componentes espaciales, peor es la calidad de los resultados. Aumentando el valor de σ_s se consigue reducir el número de falsas detecciones debidas a las vibraciones de la cámara y a los píxeles del fondo que sufren variaciones ruidosas (los valores de *Precision* aumentan). Sin embargo, los objetos móviles son peor detectados a medida que se aumenta este valor, lo cual resulta en una rápida reducción del número de píxeles móviles correctamente detectados (menor valor de *Recall*). Esto se debe a que, en las regiones en las que se encuentran los objetos móviles, cuanto mayor es el ancho espacial de los *kernels*, mayor es el número de píxeles del primer plano que están

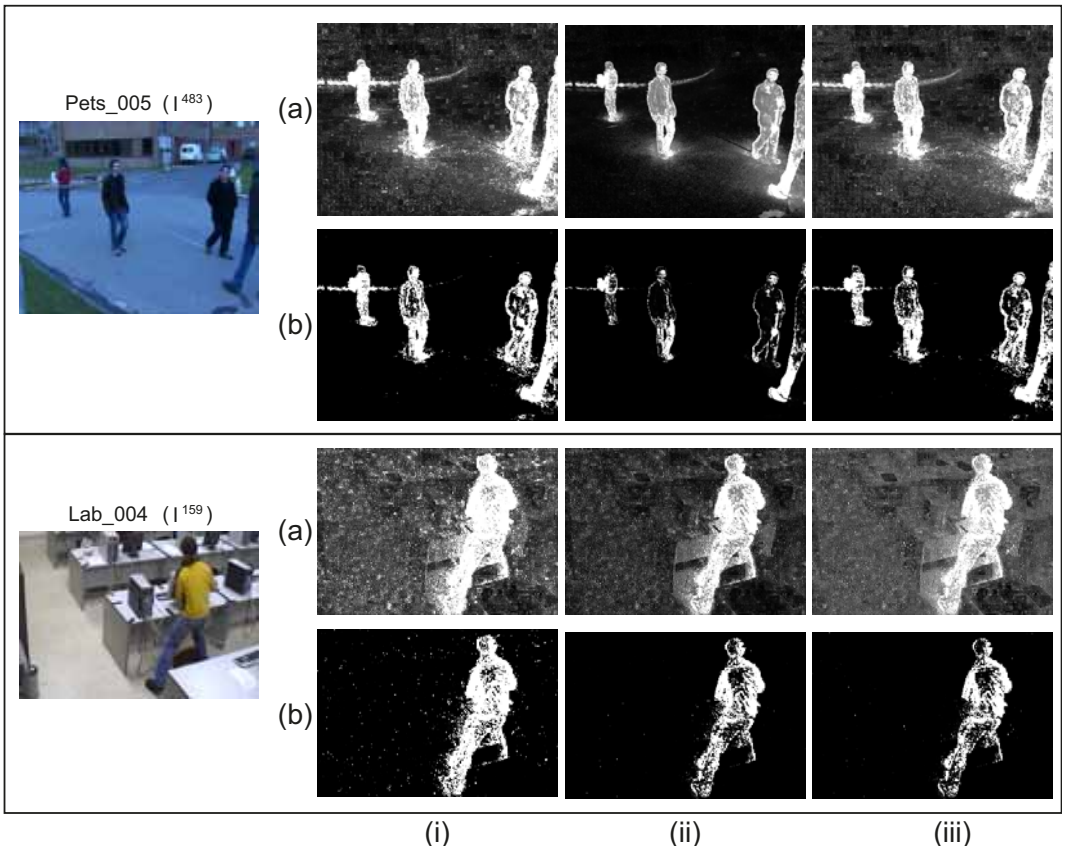


Figura 5.15: Resultados obtenidos mediante la aplicación de distintas matrices de escala en el modelado del fondo. (a) Logaritmo-negativo de la función densidad de probabilidad del fondo. (b) Detecciones finales.

siendo erróneamente utilizados como muestras de referencia para modelar el fondo, lo que da lugar una mala clasificación de estos objetos móviles. Además, esta perdida de calidad se ve incrementada si los objetos móviles son muy homogéneos.

Una solución a este problema sería modelar el fondo sin considerar las muestras correspondientes a los píxeles móviles previamente detectados. Sin embargo, de este modo tampoco se estarían considerando las variaciones sufridas por el fondo, incrementándose muy notablemente el número de falsas detecciones (Sheikh y Shah, 2005). La figura 5.14 muestra algunas detecciones que permiten comparar los resultados obtenidos en ambos casos. En la primera fila de imágenes (figura 5.14.a) se pueden ver los resultados obtenidos en el caso de utilizar muestras de referencia extraídas de todos los píxeles de las imágenes de referencia. La segunda fila de imágenes (figura 5.14.b) presenta los resultados en el caso de haberse utilizado únicamente los píxeles de las imágenes de referencia que fueron clasificados como parte del fondo. En este segundo caso se puede apreciar que, incluso utilizando anchos espaciales mayores, los objetos móviles son modelados correctamente. Sin embargo, al no actualizarse la información correspondiente a las regiones del fondo que han sufrido

variaciones, puede apreciarse que el número de falsas detecciones se ha incrementado muy notablemente.

Por lo tanto, dado que el número de falsas detecciones es considerablemente menor cuando el fondo se modela teniendo en cuenta la información de todos los píxeles (los clasificados como fondo o como primer plano), se ha decidido no descartar las muestras de los píxeles etiquetados como móviles. Además, se ha decidido hacer uso del menor ancho espacial que permita tener en cuenta píxeles vecinos (no sólo los situados en la misma posición espacial de cada píxel analizado). Por lo tanto, según el criterio establecido para considerar píxeles vecinos (aquellos que verifiquen la ecuación 5.8), dicho ancho debe tener el valor $\sigma_s = \frac{1}{3}$. De ese modo se consigue reducir las falsas detecciones debidas a los píxeles con variaciones ruidosas y a los pequeños movimientos sufridos por la cámara, sin que la cantidad de píxeles móviles no detectados sea excesivamente elevada. En los experimentos realizados se ha observado que para valores superiores de σ_s, siempre que no se aumente el número de píxeles vecinos considerados ($\sigma_s < \frac{\sqrt{2}}{3}$), la calidad de los resultados obtenidos no sufre variaciones significativas.

En último lugar, para evaluar la calidad del modelado del fondo, se han realizado algunas comparaciones entre las detecciones obtenidas mediante matrices de escala estimadas del modo descrito en la sección 5.6.1 y matrices de escala con valores fijos (Sheikh y Shah, 2005) (Zhang y Yang, 2008). En la figura 5.15 se muestran algunas de estas comparaciones, resultantes de la aplicación de dichas estrategias sobre las secuencias $Pets\_005$ y Lab_004. La segunda y tercera columnas de imágenes muestran los resultados obtenidos mediante la aplicación de *kernels* gaussianos caracterizados por matrices de escala diagonales y con idéntico valor en todas sus componentes: con $\sigma_{RGB_1} = 1$ en el caso de los mostrados en la segunda columna (figura 5.15.i), y con $\sigma_{RGB_2} = 10$ en el caso de los mostrados en la tercera columna (figura 5.15.ii). La última columna de imágenes (figura 5.15.iii) muestra los resultados obtenidos con el método propuesto. En la primera secuencia ($Pets\_005$), si se utiliza σ_{RGB_1} se obtiene una detección aceptable, pero con σ_{RGB_2} los objetos móviles no se detectan correctamente. Por otro lado, en el caso de la segunda secuencia (Lab_004), si se utiliza σ_{RGB_1} el resultado obtenido muestra demasiadas falsas detecciones, siendo mejor el obtenido con σ_{RGB_2}. Sin embargo, con el método propuesto, gracias a la estimación dinámica de las matrices de escala de los *kernels*, las detecciones obtenidas en ambas secuencias son satisfactorias (los objetos móviles se detectan adecuadamente y el número de falsas detecciones es razonablemente bajo).

5.7.2. Modelado del primer plano

En esta sección se analiza la calidad de los resultados obtenidos con la estrategia propuesta para modelar el primer plano, la cual, gracias a la etapa de actualización de las posiciones de las regiones móviles previamente detectadas y a la estimación dinámica del ancho de los *kernels* utilizados en el modelado, es capaz de mejorar los resultados de otras estrategias de modelado no paramétrico. Las detecciones mostradas en esta sección se han obtenido mediante la aplicación del clasificador bayesiano descrito en la ecuación 5.11, en el que la verosimilitud del fondo, $p(\mathbf{x}^n|\beta)$, es la resultante de la estrategia propuesta para modelar el fondo y las probabilidades a priori tienen el valor $Pr(\beta) = Pr(\phi) = \frac{1}{2}$.

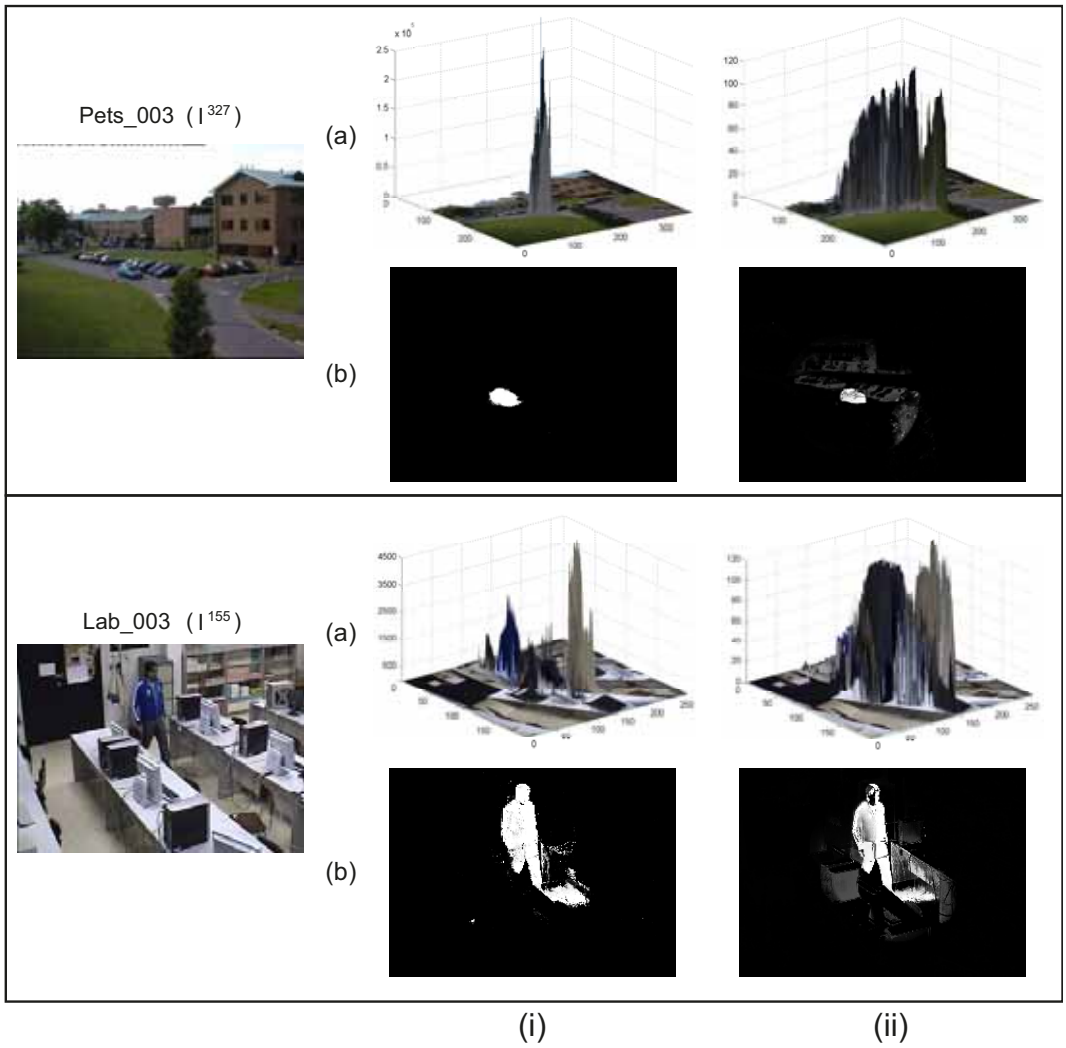


Figura 5.16: Resultados obtenidos mediante el modelado del primer plano: con la estrategia de modelado propuesta (i) y con la estrategia propuesta en (Sheikh y Shah, 2005) (ii). (a) Representación del valor de la función densidad de probabilidad estimada para el primer plano sobre las imágenes originales. (b) Detecciones finales.

Para modelar las variaciones periódicas sufridas por las regiones no estáticas del fondo de las secuencias se necesita un número de imágenes de referencia relativamente elevado. Sin embargo, dado que las regiones móviles pertenecientes al primer plano no tienden a mostrar esa periodicidad y que, además, tienden a cambiar su apariencia a medida que se desplazan, se ha decidido utilizar $N_{\phi_N} = 10$ imágenes de referencia. De esta forma, para modelar el primer plano sólo se estará teniendo en cuenta la información más reciente y, por lo tanto, probablemente más parecida a la de los objetos móviles presentes en cada instante.

En la figura 5.16 se pueden observar algunos de los resultados obtenidos sobre dos de las

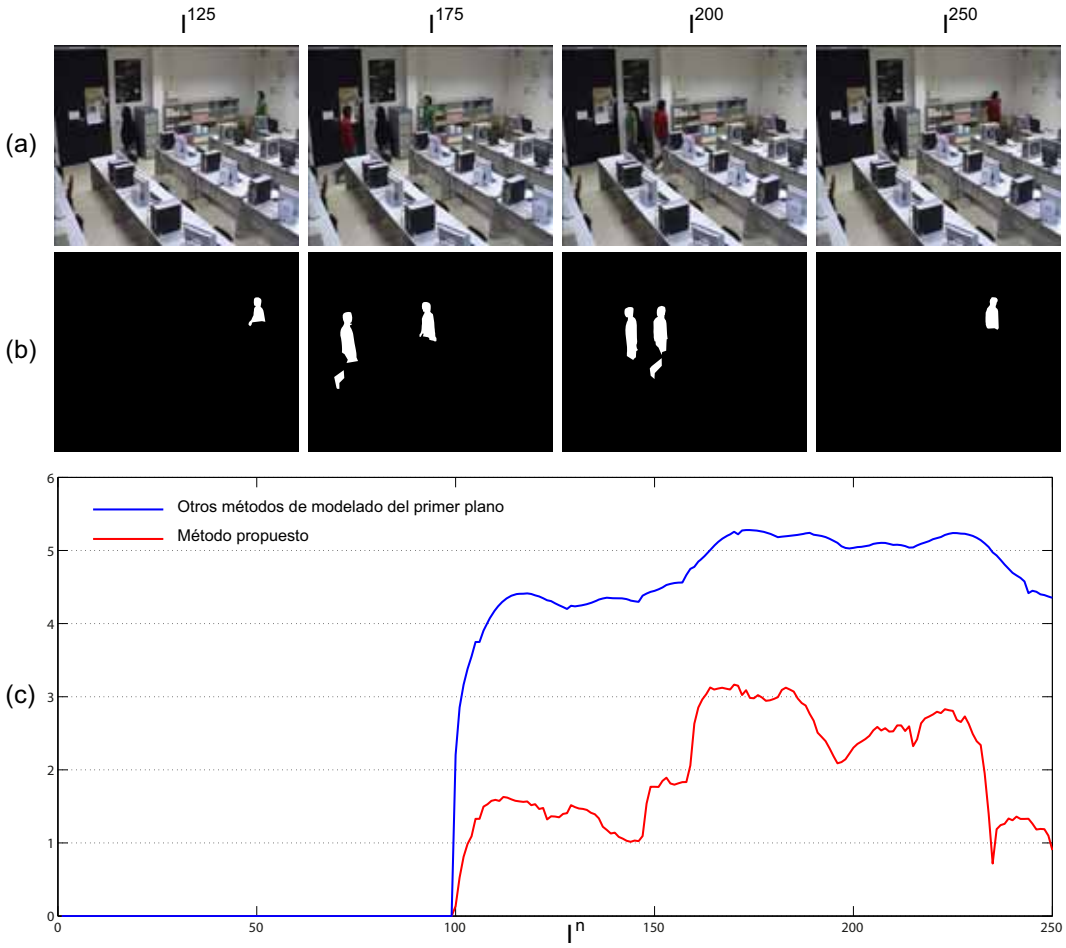


Figura 5.17: Coste computacional asociado al modelado del primer plano. (a) Imágenes representativas de la secuencia analizada. (b) Segmentaciones de referencia. (c) Logaritmo del número de *kernels* evaluados para llevar a cabo el modelado del primer plano.

secuencias de la base de datos analizada. En la segunda columna de imágenes (figura 5.16.i) se pueden ver los resultados obtenidos con el método de modelado propuesto, mientras que la tercera columna de imágenes (figura 5.16.ii) presenta los resultados obtenidos con una estrategia de modelado no paramétrico en la que ni se estima el ancho de los *kernels*, ni se actualizan las posiciones de los objetos móviles (Sheikh y Shah, 2005). Comparando los resultados proporcionados por ambos métodos se aprecia que, gracias a las mejoras propuestas para modelar el primer plano de las secuencias, la estimación de la función densidad de probabilidad asociada al primer plano es mucho más precisa, dando lugar a detecciones de mejor calidad, en las que los objetos detectados son más compactos y aparecen mejor definidos.

Además de la mejora de calidad obtenida con la estrategia propuesta, otro de los aspectos que hay que destacar es la mejora de la eficiencia computacional lograda. Esta mejora se

consigue gracias a la actualización de las regiones móviles y a la estimación dinámica del ancho de los *kernels* ya que, tal y como se ha descrito a lo largo de este capítulo, permiten reducir la cantidad de píxeles que han de ser evaluados para construir el modelo del primer plano.

En la figura 5.17 se ha representado esta mejora computacional a lo largo de una secuencia de 250 imágenes. Las primeras 97 imágenes de esta secuencia no presentan movimiento. En la imagen 98 aparece un primer objeto móvil que no desaparece hasta la imagen 234. Un segundo objeto móvil entra en escena en la imagen 156 y permanece en la misma hasta la última imagen. En la primera fila de imágenes (figura 5.17.a) se muestran algunas imágenes representativas de la secuencia analizada. En la segunda fila (figura 5.17.b) se pueden ver las detecciones de referencia de dichas imágenes. La gráfica de la parte inferior (figura 5.17.c) muestra el número de *kernels* que han requerido ser evaluados para llevar a cabo el modelado del primer plano: con las estrategias propuestas en (Sheikh y Shah, 2005) y (Zhang y Yang, 2008) en el caso de la gráfica de color azul, y con el método propuesto en el caso de la gráfica de color rojo. Mientras que en las estrategias de (Sheikh y Shah, 2005) y (Zhang y Yang, 2008) cada píxel se compara con todos los datos de primer plano previamente detectados, nuestra estrategia únicamente realiza comparaciones con los más próximos, requiriendo llevar a cabo un número de operaciones notablemente inferior (en media, unos 2 órdenes de magnitud) al de las estrategias con las que se está comparando.

Aunque con la estrategia propuesta se obtiene una más que relevante reducción del coste computacional asociado al modelado del primer plano, también se debe tener en cuenta que se han añadido otras etapas de procesamiento (seguimiento con un filtro de partículas y estimación de las matrices de escala con *Mean-Shift*) que añaden carga computacional al sistema. Sin embargo, como se verá en el siguiente capítulo, en el que se analiza detalladamente la carga computacional de cada una de estas etapas, esta carga añadida es muy inferior a la de la evaluación de los *kernels* y, por lo tanto, no suponen un incremento significativo de la carga computacional global.

5.7.3. Detecciones finales

Analizadas por separado las mejoras obtenidas con las estrategias propuestas para modelar el fondo y el primer plano, en esta sección se analizan los resultados globales del sistema de detección propuesto. En primer lugar se analiza la calidad de los resultados en función de la información de la que se hace uso: sólo el modelado del fondo, la combinación de los modelos estimados para el fondo y para el primer plano y, por último, la combinación de dichos modelos y la probabilidad a priori obtenida de las predicciones proporcionadas por el filtro de partículas utilizado para actualizar las posiciones de los objetos móviles. Algunos de estos resultados aparecen representados en la figura 5.18. La primera fila de imágenes (figura 5.18.a) muestra cuatro imágenes de distintas secuencias. La segunda fila (figura 5.18.b) muestra las detecciones de referencia correspondientes. La tercera de las filas (figura 5.18.c) presenta los resultados obtenidos en el caso de hacer uso de únicamente la información correspondiente al modelado del fondo (en las condiciones descritas en la sección 5.7.1). La cuarta fila (figura 5.18.d) muestra los resultados obtenidos tras combinar los modelados del fondo y del primer plano (en las condiciones descritas en la sección

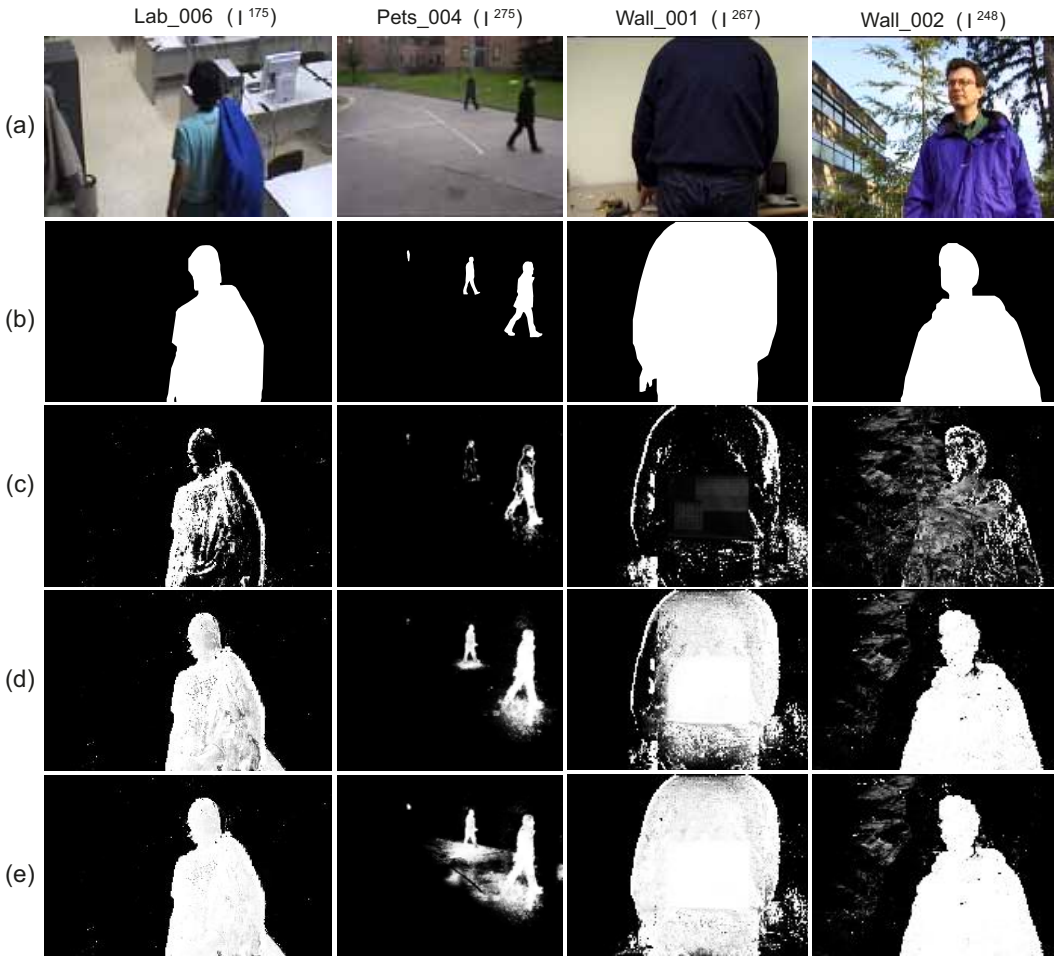


Figura 5.18: Resultados obtenidos con la estrategia de modelado no paramétrico propuesta. (a) Imágenes originales. (b) Detecciones de referencia. (c) Resultados modelando únicamente el fondo. (d) Resultados combinando los modelos de fondo y primer plano. (e) Resultados añadiendo la información a priori.

5.7.2). Por último, en la quinta fila de imágenes (figura 5.18.e) se pueden ver los resultados obtenidos tras la incorporación de la información a priori, aplicada mediante el clasificador alternativo descrito en la sección 5.4.3. Por otro lado, en la figura 5.19 se han representado los porcentajes de *Recall* y *Precision* obtenidos del análisis de los resultados de todas las secuencias de la base de datos utilizada.

Observando las imágenes de la figura 5.18 y los datos de las gráficas de la figura 5.19 se puede comprobar que utilizando únicamente el modelando del fondo se obtienen menos falsas detecciones (mayores porcentajes de *Precision*) que al combinar los modelos estimados para el fondo y para el primer plano. Sin embargo, como se puede ver en los ejemplos mostrados en la figura 5.18, en escenarios en los que el fondo y el primer plano no se diferencian lo suficiente entre sí, o en situaciones en las que los objetos móviles se quedan parados breve-

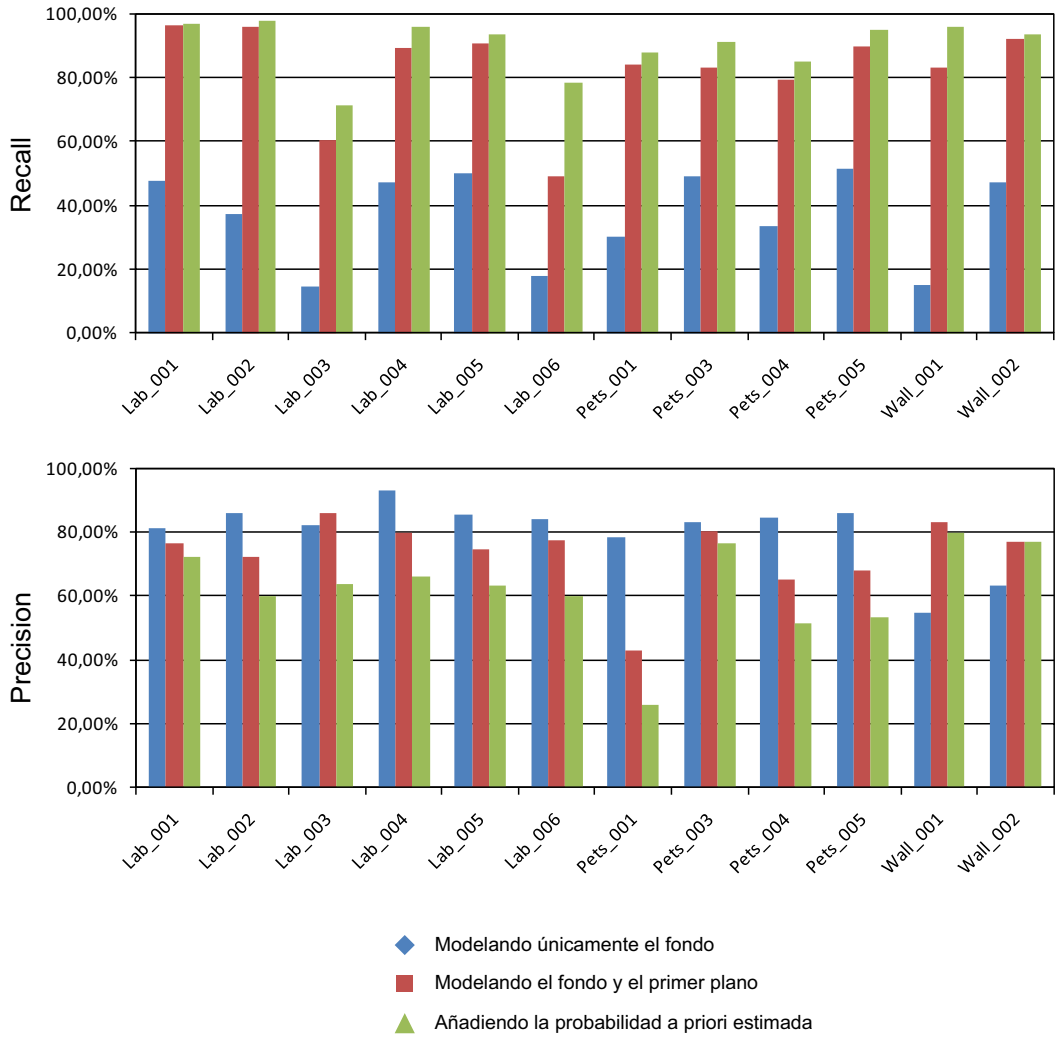


Figura 5.19: Porcentajes de *Recall* y *Precision* en función de la información que se utiliza en la detección.

mente (por ejemplo, en los instantes analizados en las dos últimas columnas de imágenes de la figura 5.18), el número de píxeles móviles no detectados aumenta muy notablemente si únicamente se utiliza el modelado correspondiente al fondo (se obtienen valores bajos de *Recall*). Este elevado número de píxeles móviles no detectados se ve significativamente reducido cuando además de utilizarse el modelo estimado para el fondo, también se hace uso del modelado del primer plano, obteniéndose de este modo mayores valores de *Recall*. Estos porcentajes de *Recall* se incrementan todavía más si, además de utilizarse ambos modelados, se incorpora la información a priori obtenida del modo descrito en la sección 5.5.2. Sin embargo, hay que apreciar que en la mayor parte de las secuencias los objetos móviles provocan sombras y reflejos a su paso y, al igual que los objetos móviles, éstas también son mejor detectadas al utilizar el modelado del primer plano y las probabilidades

	Mezcla de gaussianas			**Modelado de (Sheikh y Shah, 2005)**			**Método propuesto**		
	Recall	Precision	**F**	Recall	Precision	**F**	Recall	Precision	**F**
Lab_001	73,19	83,92	78,19	88,14	47,06	61,36	96,88	72,06	82,64
Lab_002	76,39	84,37	80,18	87,20	71,48	78,56	97,68	60,08	74,40
Lab_003	77,56	65,72	71,15	62,72	77,58	69,36	71,37	63,73	67,34
Lab_004	83,60	69,24	75,75	72,66	90,56	80,63	95,96	66,35	78,46
Lab_005	69,03	84,09	75,82	88,79	49,31	63,41	93,59	63,18	75,43
Lab_006	91,20	58,98	71,63	44,36	74,81	55,70	78,21	60,09	67,96
Pets_001	87,54	71,77	78,87	76,88	33,05	46,23	87,93	26,12	40,27
Pets_002	100	0	0	100	0	0	100	0	0
Pets_003	81,49	69,21	74,85	48,99	86,58	62,57	91,24	76,44	83,19
Pets_004	74,37	81,94	77,97	58,14	91,09	70,98	85,09	51,29	64,01
Pets_005	82,34	90,28	86,13	64,28	95,68	76,90	94,97	53,32	68,30
Wall_001	95,49	43,83	60,09	84,30	99,00	91,06	95,90	79,61	87,00
Wall_002	95,55	6,72	12,56	90,45	93,65	92,03	93,71	77,05	84,57
Promedio	86,78	20,84	33,60	72,88	84,97	78,46	91,72	59,16	71,93

Tabla 5.2: Resultados cualitativos obtenidos con la estrategia propuesta, comparados con los obtenidos mediante otras técnicas de detección.

a priori (en el ejemplo mostrado en la segunda columna de imágenes de la figura 5.18 puede apreciarse claramente la influencia de las sombras en las detecciones). Consecuentemente, utilizando el modelado del primer plano y las probabilidades a priori aumenta el número de falsas detecciones y se reducen los porcentajes de *Precision* (en mayor medida para las secuencias con mayor contenido de sombras y reflejos).

En último lugar se ha llevado a cabo una comparación cuantitativa y cualitativa entre los resultados obtenidos con la estrategia descrita en este capítulo y los obtenidos mediante otras técnicas de detección de objetos móviles: el método paramétrico presentado en el capítulo 4 y el método no paramétrico propuesto en (Sheikh y Shah, 2005). El resumen de los resultados cuantitativos (en términos de porcentajes de *Recall*, *Perecision* y F) se muestra en la tabla 5.2 y en las figuras 5.20, 5.21 y 5.22 se han representado algunas detecciones que permiten analizar situaciones relevantes. En la primera fila de imágenes de estas figuras se muestran las imágenes originales analizadas. En la segunda fila se han representado las detecciones de referencia de dichas imágenes. En la tercera fila se pueden ver los resultados obtenidos mediante el método de mezcla de gaussianas descrito en el capítulo 4. En la cuarta fila de imágenes se muestran los resultados obtenidos con las estrategias de modelado no paramétrico propuestas en (Sheikh y Shah, 2005). Por último, en la quinta fila de imágenes se pueden observar los resultados obtenidos con la estrategia propuesta en el presente capítulo.

Prestando atención a los resultados mostrados en las tres figuras y a los datos que contiene la tabla 5.2 se puede comprobar que, generalmente, los métodos basados en técnicas de modelado no paramétrico (especialmente el método propuesto) dan lugar a detecciones

Figura 5.20: Análisis cualitativo de los resultados obtenidos con la estrategia propuesta y con otros métodos de detección de objetos. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con el método de mezcla de gaussianas presentado en el capítulo 3. (d) Detecciones obtenidas con las estrategias de modelado no paramétrico de (Sheikh y Shah, 2005). (e) Detecciones obtenidas con el método no paramétrico propuesto en este capítulo.

con menor cantidad de píxeles móviles no detectados (mayores porcentajes de *Recall*) y, además, en escenarios en los que el fondo sufre variaciones multimodales, reducen muy notablemente el número de falsas detecciones (mayores porcentajes de *Precision*). Además, esta reducción del número de falsas detecciones es especialmente significativa en el caso de secuencias con fondos muy dinámicos, como es el caso de $Wall\_001$ y $Wall\_002$. Las dos últimas columnas de la figura 5.20 muestran dos ejemplos de las detecciones obtenidas en estas dos secuencias, en los que se puede apreciar que los métodos paramétricos no son capaces de modelar adecuadamente las variaciones sufridas por el fondo.

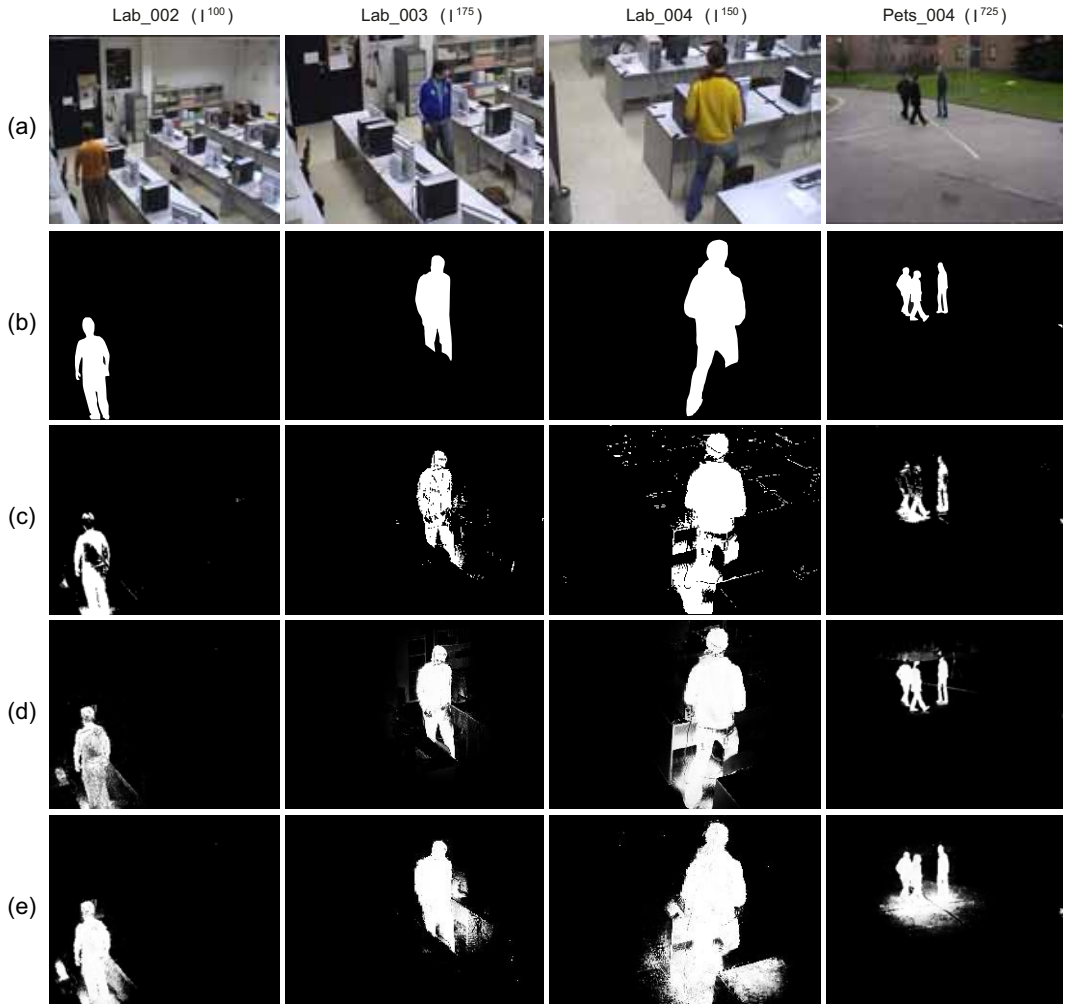


Figura 5.21: Influencia de las sombras en las detecciones. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con el método de mezcla de gaussianas presentado en el capítulo 3. (d) Detecciones obtenidas con las estrategias de modelado no paramétrico de (Sheikh y Shah, 2005). (e) Detecciones obtenidas con el método propuesto en este capítulo.

Sin embargo, si se presta atención a los datos mostrados en la tabla 5.2, una de las cosas que más llama la atención es que, en muchas de las secuencias analizadas, los porcentajes de *Precision* resultantes de la aplicación de las estrategias basadas en el modelado no paramétrico son peores que los obtenidos con el método basado en la mezcla de gaussianas. Esto se debe a que con los métodos de modelado no paramétrico, además de detectarse mejor los objetos móviles, también se detectan mejor las sombras y los reflejos que provocan en sus desplazamientos, incrementándose significativamente el número de falsas detecciones en dichas situaciones. En la figura 5.21 se pueden ver cuatro ejemplos en los que se muestran las detecciones sobre cuatro escenarios en los que los objetos móviles son causantes de sombras y

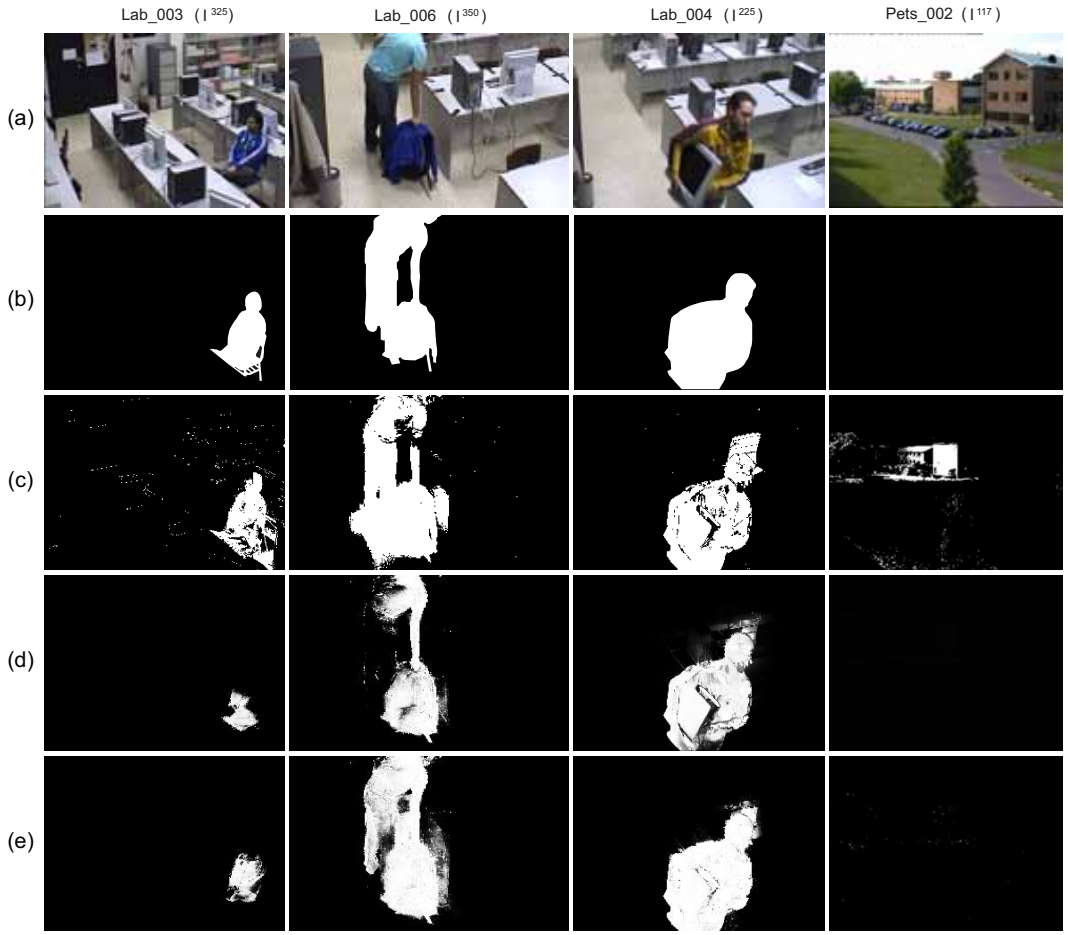


Figura 5.22: Análisis de la velocidad de actualización del fondo. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con el método de mezcla de gaussianas presentado en el capítulo 3. (d) Detecciones obtenidas con las estrategias de modelado no paramétrico de (Sheikh y Shah, 2005). (e) Detecciones obtenidas con el método propuesto en este capítulo.

reflejos. Como muestran dichos ejemplos, las detecciones obtenidas con los métodos basados en el modelado no paramétrico constan de menor número de píxeles móviles no detectados y menor cantidad de falsas detecciones debidas a las variaciones multimodales del fondo. Sin embargo, las sombras y reflejos que aparecen alrededor de los objetos también son detectadas como parte del primer plano. Hay que destacar que al igual que el método propuesto mejora la detección de los objetos móviles frente a los métodos de modelado propuestos en (Sheikh y Shah, 2005), también mejora la detección de sus sombras y, consecuentemente, da lugar a los menores porcentajes de *Precision*.

Por otro lado, también hay que mencionar que las matrices de escala que se utilizan en (Sheikh y Shah, 2005) definen *kernels* con un ancho lo suficientemente grande como para

obtener resultados con un número muy bajo de falsas detecciones debidas a las variaciones multimodales del fondo pero, a su vez, dicho ancho hace que en muchos escenarios los objetos móviles no sean detectados adecuadamente. Además, debido a que hacen uso de anchos espaciales grandes ($\sigma_s = 25$), cuando los objetos móviles se encuentran en zonas con colores similares a las regiones del fondo, dan lugar a la aparición de falsas detecciones alrededor de los objetos móviles (se pueden apreciar estas situaciones en los ejemplos mostrados en las dos primeras columnas de la figura 5.20).

Para terminar se debe hacer mención a los casos en los que los porcentajes de *Recall* obtenidos con el método propuesto son peores que los obtenidos con el método paramétrico descrito en el capítulo 4 (secuencias Lab_003, Lab_006 y $Pets\_001$). Esta pérdida de calidad se debe a que dichas secuencias contienen objetos que, en algún momento a lo largo de las mismas, se quedan parados y, al contrario que en la estrategia basada en gaussianas, en la que se es posible decidir el tiempo aproximado en el que dichos objetos pasen a formar parte del fondo, los métodos de modelado no paramétrico actualizan el fondo mucho más rápido. Por otro lado, esta rápida actualización del fondo, al igual que es perjudicial en situaciones en las que los objetos móviles se quedan parados, es beneficiosa en escenarios en los que el fondo sufre variaciones rápidas (sustracción de objetos, cambios de iluminación, etc.). En la figura 5.22 se muestran los resultados correspondientes a cuatro situaciones en las que la velocidad de actualización del fondo influye enormemente en la calidad de las detecciones. Las imágenes de las dos primeras columnas se corresponden con situaciones en las que dos objetos móviles se quedan parados. Las detecciones correspondientes a estas situaciones muestran cómo la rápida actualización del fondo reduce la calidad de los métodos basados en el modelado no paramétrico (algo menos en el caso de la estrategia propuesta). Las imágenes mostradas en las dos últimas columnas se corresponden, respectivamente, con un escenario en el que se ha sustraido un elemento del fondo (un monitor de ordenador) y con una escena en la que se está produciendo un cambio de iluminación que afecta a gran parte de la misma. En ambos casos se puede apreciar que la rápida actualización del fondo en las estrategias no paramétricas mejora la calidad de los resultados, reduciendo notablemente el número de falsas detecciones.

Finalmente, en el caso de la secuencia $Pets\_002$, al igual que se ha hecho en el capítulo 4 y por los mismos motivos, para determinar la calidad de las detecciones obtenidas se ha analizado la cantidad de falsas detecciones resultantes de la aplicación de las tres estrategias comparadas. Mientras que la estrategia basada en la mezcla de gaussianas ha dado lugar a 2160138 falsas detecciones, con las estrategias de modelado de (Sheikh y Shah, 2005) y con la estrategia propuesta en este capítulo dicha cantidad se ha reducido, respectivamente, a 51 y a 58193 falsas detecciones. El motivo de que con el método propuesto en (Sheikh y Shah, 2005) se obtengan menos falsas detecciones que con nuestro método es que, por defecto, como ya se ha mencionado anteriormente, en (Sheikh y Shah, 2005) se utilizan *kernels* muy anchos: se prima la supresión de falsas detecciones, a costa de un mayor número de píxeles móviles no detectados.

5.8. Conclusiones

En este capítulo se ha descrito una estrategia, basada en técnicas de modelado no paramétrico, que permite la detección eficiente de los objetos móviles presentes en secuencias de vídeo.

Para reducir la cantidad de píxeles móviles no detectados en situaciones en las que el contenido móvil de las secuencias es muy parecido al de las regiones del fondo, se ha hecho uso de un clasificador bayesiano que combina un modelo estimado para el fondo con otro estimado para el primer plano. Para obtener la estimación de ambos modelos se han utilizado muestras de referencia extraídas de los píxeles de imágenes previas, cuyas coordenadas los sitúan cerca de la posición espacial del píxel de la imagen actual que referencian. De ese modo se ha conseguido reducir la cantidad de falsas detecciones debidas a las vibraciones de la cámara y a las regiones no estáticas del fondo.

Además, para mejorar la calidad de cada uno de los modelados, se han propuesto dos estrategias para la estimación dinámica de las matrices de escala que determinan los anchos de los *kernels* de los que hacen uso: en el caso del modelado del fondo se ha utilizado una estrategia basada en el análisis estadístico de los valores absolutos de las diferencias entre píxeles situados en la misma posición espacial de imágenes consecutivas; mientras que para el primer plano se ha diseñado una estrategia, basada en *Mean-Shift*, que permite la agrupación de los píxeles móviles previamente detectados en regiones homogéneas.

Por otro lado, para mejorar la calidad del modelado del primer plano se ha aplicado una estrategia, basada en la actualización de las regiones móviles previamente detectadas, que permite obtener resultados más precisos y que, además, consigue una importante reducción del coste computacional asociado a dicho modelado. Esta actualización se ha llevado a cabo mediante la utilización de un filtro de partículas diseñado para seguir un número variable de regiones móviles. Adicionalmente, de los resultados proporcionados por dicho filtro se ha obtenido información relativa a la localización más probable de los objetos móviles en el futuro que, combinada con los modelos estimados para el fondo y para el primer plano, ha hecho posible obtener una mejora adicional de los resultados.

Para evaluar la calidad los resultados obtenidos se han analizado numerosas secuencias con contenido crítico para la detección de objetos móviles (fondos dinámicos, cambios de iluminación, objetos móviles parecidos al fondo, objetos móviles que se quedan parados, etc.) y los resultados obtenidos se han comparado con los resultantes de la aplicación de otras estrategias basadas tanto en métodos paramétricos como en métodos no paramétricos. Con los análisis y las comparaciones efectuadas se ha comprobado que la estrategia propuesta proporciona resultados de gran calidad, reduciendo muy notablemente la cantidad de falsas detecciones en escenarios con fondos multimodales y dando lugar a detecciones con muy poca cantidad de píxeles móviles no detectados. En general, los resultados obtenidos con la estrategia propuesta han mostrado ser mejores que los obtenidos con las estrategias con las que se ha comparado.

Además, gracias a la estimación dinámica de los anchos de los *kernels* se ha conseguido una apreciable reducción de la cantidad de *kernels* que han de ser evaluados para llevar a cabo el modelado del primer plano y, consecuentemente, se ha logrado una apreciable reducción del coste computacional asociado a dicho modelado.

Capítulo 6

Reducción del coste computacional y supresión de sombras y reflejos

De nada sirve al hombre lamentarse de los tiempos en que vive. Lo único bueno que puede hacer es intentar mejorarlos.

Thomas Carlyle (1795-1881),
historiador, crítico social y ensayista escocés.

Resumen: Los métodos de detección de objetos móviles basados en técnicas de modelado no paramétrico, tal y como se ha mostrado en el capítulo anterior, son capaces de obtener resultados de gran calidad, evitando las falsas detecciones debidas a las variaciones multimodales del fondo de las secuencias y detectando la mayor parte de los píxeles móviles. Sin embargo, estos métodos tienen asociado un elevado coste computacional que dificulta su utilización en aplicaciones que requieren trabajar a gran velocidad. Además, aunque eliminan las falsas detecciones debidas a las variaciones del fondo, no son capaces de descartar aquellas debidas a las sombras y reflejos provocados por los objetos móviles. Es por eso que se ha desarrollado la estrategia que se describe en este capítulo, la cual, tomando como base las estrategias de modelado descritas en el capítulo 5, permite reducir su coste computacional y mejorar notablemente sus resultados. Por un lado, para reducir el número de falsas detecciones debidas a las sombras y los reflejos que originan los objetos móviles, se propone la utilización de un novedoso y efectivo conjunto de características de apariencia para llevar a cabo los modelados del fondo y del primer plano. Por otro lado, para mejorar la eficiencia computacional de la estrategia, se propone la utilización de la información espacial de los píxeles únicamente en el modelado del primer plano y se aplica una estrategia que, a partir de pequeñas regiones aleatoriamente repartidas por las imágenes y de la predicción del filtro de partículas utilizado para actualizar las posiciones de los objetos previamente detectados, permite obtener máscaras de regiones de interés que determinan qué píxeles han de ser analizados y cuáles no.

6.1. Introducción

Tal y como se ha visto en el capítulo anterior, las técnicas de detección de objetos basadas en el modelado no paramétrico son capaces de obtener detecciones de gran calidad, mejorando los resultados obtenidos mediante técnicas de detección paramétricas como la presentada en el capítulo 4.

Estas técnicas, para obtener resultados de calidad en secuencias con fondos dinámicos y en las que los objetos móviles son parecidos al fondo, modelan tanto el fondo como el primer plano. De esa forma son capaces de conseguir detecciones robustas que mejoran la calidad obtenida por otras estrategias: obtienen menor número de píxeles móviles no detectados y reducen muy notablemente la cantidad de falsas detecciones debidas a las variaciones multimodales de los fondos.

Sin embargo, como se ha podido comprobar en el análisis de los resultados llevado a cabo en el capítulo 5, las estrategias de detección que modelan tanto el fondo como el primer plano tienen algunos inconvenientes que han de ser tenidos en consideración. Entre estos inconvenientes destacan: por un lado, las falsas detecciones debidas a las sombras y a los reflejos que provocan los objetos móviles en sus desplazamientos; y por otro lado, su elevado coste computacional asociado.

En este capítulo se describe una estrategia de detección de objetos móviles, basada en la presentada en el capítulo 5, que mejora muy notablemente su eficiencia computacional y que, además, elimina la mayor parte de las falsas detecciones debidas a las sombras y los reflejos provocados por los objetos móviles.

Por un lado, para llevar a cabo los modelados del fondo y del primer plano se utiliza una innovadora combinación de color normalizado y gradientes. De ese modo se consigue reducir la influencia de las sombras y reflejos en las detecciones, a la vez que se consigue mejorar la cantidad de detecciones correctas en escenarios en los que las componentes RGB no permiten discriminar adecuadamente entre fondo y primer plano.

Por otro lado, dado que el mayor coste computacional se debe a las etapas correspondientes al modelado del fondo (estimación de las matrices de escala de los *kernels* y evaluación de dichos *kernels*), para reducir dicho coste se propone limitar el número de muestras de referencia asociadas a cada píxel a aquellas situadas en su misma posición espacial. De ese modo la información espacial de los píxeles se utilizará únicamente en el modelado del primer plano en el que, dada la naturaleza móvil de los objetos del primer plano, no se puede eliminar dicha información espacial.

Además, para conseguir una mejora computacional todavía mayor, se propone la estimación de máscaras de regiones de interés, obtenidas a partir de la predicción proporcionada por el filtro de partículas utilizado para actualizar las posiciones de los objetos móviles y de pequeñas regiones aleatoriamente repartidas por las imágenes. Estas máscaras permiten determinar qué zonas de las imágenes deben ser analizadas en cada instante y cuáles de ellas no es necesario analizar y, por lo tanto, evitan el análisis de un gran porcentaje de los píxeles de las imágenes.

En primer lugar, en la sección 6.2 se describe el conjunto de características de apariencia propuesto para evitar las falsas detecciones debidas a sombras y reflejos. A continuación,

en la sección 6.3 se describen las estrategias de modelado del fondo y del primer plano, el clasificador bayesiano mediante el que se combinan ambos modelados y el modo de obtención de las probabilidades a priori utilizadas en dicho clasificador. Acto seguido, en la sección 6.4 se analiza el modo de generación de las máscaras de regiones de interés utilizadas para determinar qué píxeles deben analizarse en cada imagen. Por último, en las secciones 6.5 y 6.6 se presentan, respectivamente, los resultados y las conclusiones del capítulo.

6.2. Supresión de sombras y reflejos

Trabajando en un espacio típico de componentes RGB, la aparición de sobras y reflejos provocados por los objetos móviles es muy común. Normalmente, estas sombras y reflejos son erróneamente clasificados como parte del primer plano de las secuencias, reduciendo la calidad de las detecciones obtenidas y dificultando la posterior aplicación de otras fases de análisis (Elhabian et al., 2008).

Para solucionar estos problemas, discriminando correctamente entre los objetos móviles y sus sombras o reflejos, en vez de utilizar las componentes RGB es posible utilizar las componentes de color normalizadas. Bajo ciertas condiciones, las componentes de color normalizadas resultan invariantes a los cambios de iluminación (Amato et al., 2011), reduciendo por lo tanto la influencia de las sombras y los reflejos en las detecciones. Es por eso que en algunos trabajos (Lin y Lee, 1997) se ha propuesto hacer uso de dichas componentes, las cuales se definen como:

$$Rn = \frac{R}{R+G+B} \qquad Gn = \frac{G}{R+G+B} \tag{6.1}$$

donde R, G y B son las componentes RGB de los píxeles.

No obstante, si se utilizan estas componentes, las regiones caracterizadas principalmente por información de luminancia (regiones grises) no se modelan adecuadamente. Para solucionar este problema, algunos autores (Elgammal et al., 2002) (Mittal y Paragios, 2004) añaden a este par de componentes normalizadas la información de saturación:

$$S = R + G + B \tag{6.2}$$

Sin embargo, esta combinación no proporciona resultados totalmente satisfactorios ya que, aunque en menor medida que en el caso de utilizar las componentes RGB, las sombras y los reflejos vuelven a estar presentes en las detecciones.

Para resolver estas limitaciones, detectando correctamente tanto el color de las regiones móviles como su información de luminancia, nosotros proponemos utilizar una novedosa y efectiva combinación de características, compuesta por las componentes de color normalizadas y por el módulo del gradiente de la saturación, $|\nabla S|$. Por lo tanto, el vector de componentes de apariencia que proponemos es de la forma:

$$\mathbf{c} = (Rn, Gn, |\nabla S|)^T \tag{6.3}$$

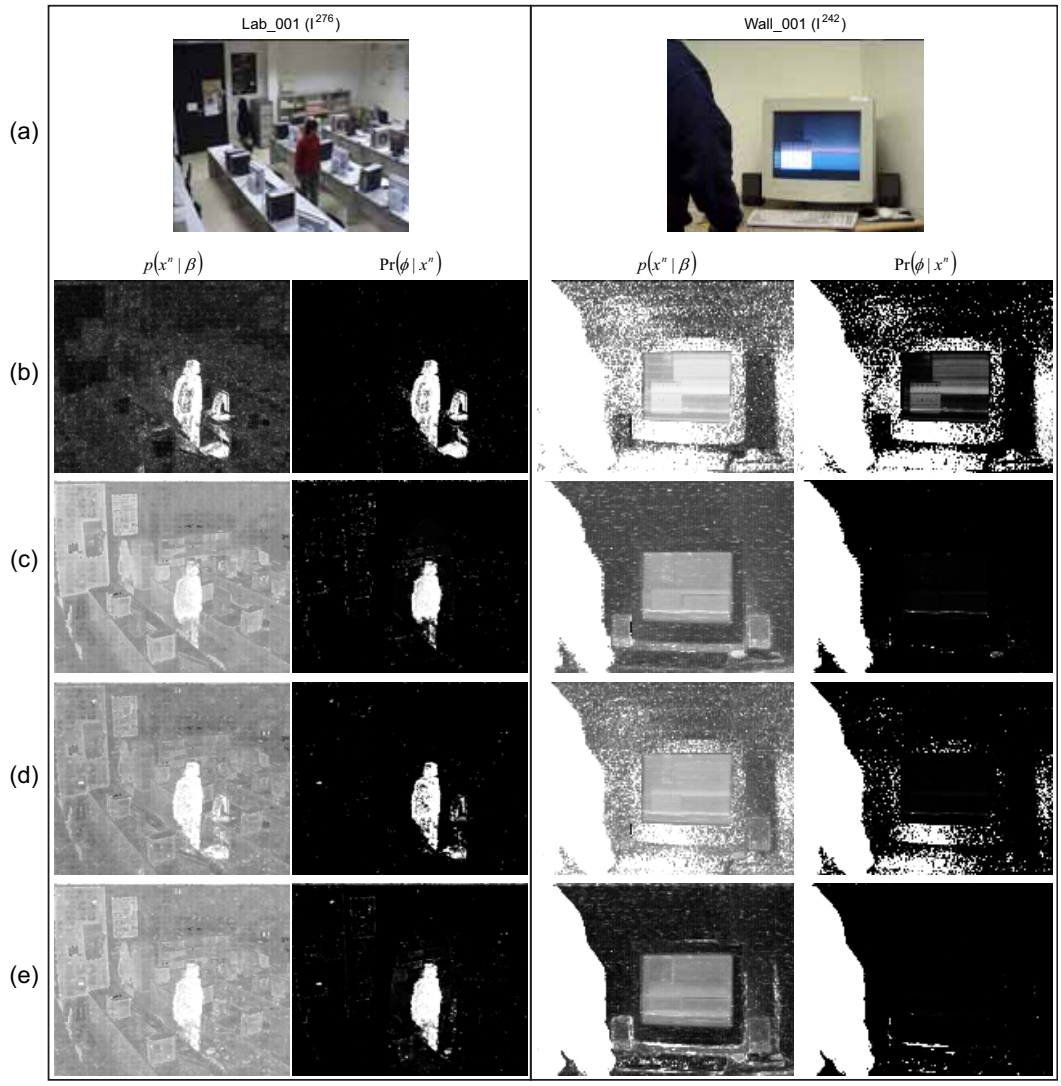


Figura 6.1: Detecciones obtenidas mediante distintos conjuntos de componentes de apariencia de los píxeles. (a) Imágenes originales. (b) Resultados obtenidos con las componentes de color RGB. (c) Resultados obtenidos con las componentes de color normalizadas. (d) Resultados obtenidos con las componentes de color normalizadas y la saturación. (e) Resultados obtenidos con el conjunto de componentes propuesto.

Mientras que los sistemas basados en el análisis del color son susceptibles a las variaciones debidas a las sombras y a los cambios de iluminación, los gradientes son menos afectados por estas variaciones (Mittal y Paragios, 2004) y pueden ser combinados de forma muy efectiva con la información de color. Además, en las situaciones en las que el fondo y el primer plano tienen colores muy similares, la utilización de los gradientes ayuda a discriminar entre ellos. Otra ventaja de los gradientes es que, al depender de los valores de los píxeles en un

Figura 6.2: Supresión de falsas detecciones mediante filtrados morfológicos. (a) Imágenes originales. (b) Detecciones obtenidas. (c) Detecciones tras aplicar un filtrado morfológico.

entorno espacial, proporcionan información de más alto nivel que las componentes de color. Por todas estas razones, la combinación de color y gradientes resulta muy recomendable para llevar a cabo los modelados del fondo y del primer plano.

En la figura 6.1 se muestran algunos de los resultados obtenidos mediante la estrategia de detección presentada en el capítulo 5, haciendo uso de distintos conjuntos de componentes de apariencia de los píxeles. Para cada caso analizado se han representado: el logaritmo-negativo de la función densidad de probabilidad estimada para el fondo, $p(\mathbf{x}^n|\beta)$, y las detecciones finales obtenidas, $Pr(\phi|\mathbf{x}^n)$. Utilizando las componentes RGB de los píxeles (figura 6.1.b), se aprecia la gran cantidad de falsas detecciones debidas a la clasificación errónea como parte del primer plano de las sombras y los reflejos causados por los objetos móviles. Por el contrario, si las detecciones se realizan haciendo uso de las componentes de color normalizadas (figura 6.1.c), al ser estas más invariantes a los cambios de luminosidad, las sombras y los reflejos son modelados correctamente y, por lo tanto, las detecciones son mucho mejores. Sin embargo, esta invarianza también hace que las regiones grises de los objetos móviles no sean adecuadamente modeladas (por ejemplo, los pantalones del objeto móvil que aparece en el ejemplo de la izquierda de la figura), y esto resulta en un aumento del número de píxeles móviles no detectados. Si, para evitar estos errores en el modelado de los objetos móviles, se añade la información aportada por la saturación (figura 6.1.d), las sombras y los reflejos vuelven a ser clasificados erróneamente (aunque en menor grado

que cuando se utilizan las componentes RGB). Sin embargo, si se utiliza el conjunto de componentes propuesto (figura 6.1.e), tanto las sombras y reflejos como las regiones grises de los objetos móviles son modeladas correctamente, obteniéndose mejores detecciones que en cualquiera de los casos anteriores.

6.3. Modelado no paramétrico del fondo y del primer plano

Al modelar el fondo de las secuencias a partir de muestras de referencia extraídas de los píxeles de imágenes previas dentro de un margen espacial entorno a la posición de cada píxel de la imagen actual, se reduce la cantidad de falsas detecciones (tanto las debidas a las vibraciones de la cámara como las resultantes de las variaciones ruidosas de algunos píxeles del fondo). En un modelado de este tipo, para obtener la verosimilitud asociada a cada uno de los píxeles de la imagen actual hay que evaluar un gran número de *kernels* multidimensionales (mayor a medida que se tiene en cuenta más información espacial), lo cual resulta en un coste computacional demasiado elevado como para pensar en utilizar este tipo de estrategias en aplicaciones que requieran trabajar en tiempo real. Además, generalmente, gran parte de las falsas detecciones debidas a los efectos previamente mencionados también pueden ser descartadas mediante la aplicación de distintos filtros morfológicos (Salembier y Wilkinson, 2009), lo cual reduce todavía más la necesidad de utilizar información espacial en el modelado del fondo. En la figura 6.2 se han representado tres ejemplos en los que, aplicando un filtrado de tipo apertura (Gao et al., 2008), se ha eliminado la mayor parte de las falsas detecciones obtenidas tras llevar a cabo un modelado del fondo que no considera información espacial.

Por otro lado, debido que los objetos del primer plano modifican notablemente sus posiciones en periodos breves de tiempo, el modelado correspondiente al primer plano sí que requiere que los píxeles de referencia de los que hace uso sean los comprendidos dentro de un margen espacial.

Atendiendo a estas razones, en esta sección se propone una estrategia, basada en el modelado no paramétrico del fondo y el primer plano, en la que la información espacial de los píxeles de referencia se utiliza únicamente en el modelado correspondiente al primer plano. De ese modo, el coste computacional asociado al modelado del fondo se reduce de forma muy significativa.

Modelado del fondo Considérese un píxel p^n en el instante n. El modelado del fondo, a diferencia que en el propuesto en el capítulo 5, se va a llevar a cabo a partir de muestras de referencia situadas en la misma posición espacial de dicho píxel y, por lo tanto, únicamente será necesario tener en cuenta sus componentes de apariencia, pudiéndose descartar las componentes espaciales descritas en el capítulo 5. Además, para reducir la influencia de las sombras en el modelado, dichas componentes de apariencia serán las descritas en la sección 6.2 (componentes de color normalizadas y módulo del gradiente de la saturación). Por lo tanto, de acuerdo con el modelo de estimación descrito en el capítulo 5, utilizando *kernels* gaussianos con matrices de escala diagonales, la verosimilitud de p^n de pertenecer al fondo

de la secuencia, β, se puede estimar como:

$$p(\mathbf{c}^n|\beta) = \frac{1}{N_\beta(2\pi)^{\frac{3}{2}}} \sum_{i=1}^{N_\beta} \prod_{j=1}^{3} \frac{1}{(\Sigma_\beta(j,j))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{(\mathbf{c}^n(j) - \mathbf{c}^i_\beta(j))^2}{\Sigma_\beta(j,j)}\right) \tag{6.4}$$

donde $\mathbf{c}^n = (Rn^n, Gn^n, |\nabla S|^n)^T$ es el vector de características de apariencia de p^n, $\{\mathbf{c}^i_\beta\}_{i=1}^{N_\beta}$ son las muestras de referencia extraídas de las N_β imágenes previas y Σ_β es la matriz de escala que determina el ancho de los *kernels*. Dicha matriz de calcula del modo descrito en la sección 5.6.1 del capítulo 5.

Modelado del primer plano Mientras que el modelado del fondo se realiza sin hacer uso de ninguna información espacial de los píxeles, como ya se ha dicho anteriormente, el modelado del primer plano sí que utiliza esta información y, por lo tanto, se puede llevar a cabo del modo descrito en la sección 5.4.2 del capítulo 5. Así, la verosimilitud de que el píxel p^n forme parte del primer plano de la secuencia, ϕ, se puede calcular como la mezcla de una función uniforme, γ, y la estimación de una función densidad de probabilidad con *kernels*, a partir de un conjunto de N_ϕ muestras de referencia, $\{\mathbf{x}^i_\phi = ((\mathbf{c}^i_\phi)^T, (\mathbf{s}^i_\phi)^T)^T\}_{i=1}^{N_\phi}$, extraídas en este caso de los píxeles pertenecientes a las N_{ϕ_N} imágenes previas a la actual, dentro de un entorno espacial definido del modo descrito en la sección 5.4.2 del capítulo 5, donde $\mathbf{c}^i_\phi$ es el vector de características de apariencia (componentes de color normalizadas y módulo del gradiente de la saturación) de los píxeles de referencia y $\mathbf{s}^i_\phi$ es el vector que contiene sus componentes espaciales:

$$p(\mathbf{x}^n|\phi) = \alpha\gamma + \frac{(1-\alpha)}{N_\phi(2\pi)^{\frac{5}{2}}} \sum_{i=1}^{N_\phi} \prod_{j=1}^{5} \frac{1}{(\Sigma_\phi(j,j))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}^n(j) - \mathbf{x}^i_\phi(j))^2}{\Sigma_\phi(j,j)}\right) \tag{6.5}$$

donde $\mathbf{x}^n = ((\mathbf{c}^n)^T, (\mathbf{s}^n)^T)^T$ es el vector de características (apariencia y posición espacial) extraídas de p^n, α es el factor de mezcla de las dos funciones y Σ_ϕ es la matriz de escala de los *kernels*, estimada del modo descrito en el capítulo 5.

Clasificador bayesiano De forma similar a la estrategia descrita en el capítulo 5, una vez obtenidos los modelos del fondo y del primer plano, la probabilidad de que cada píxel pertenezca a una u otra clase se puede obtener utilizando el clasificador bayesiano alternativo propuesto en la sección 5.4.3:

$$Pr(\phi|\mathbf{x}^n) = \frac{Pr(\phi|\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n,\phi)}{Pr(\phi|\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n,\phi) + Pr(\beta|\mathbf{s}^n)p(\mathbf{c}^n|\beta)} \tag{6.6}$$

donde $p(\mathbf{c}^n|\mathbf{s}^n,\beta)$, al no utilizarse información espacial para modelar el fondo, se ha sustituido por $p(\mathbf{c}^n|\beta)$, $p(\mathbf{c}^n|\mathbf{s}^n,\phi)$ es la probabilidad condicionada en el espacio que se definió en la ecuación 5.13 y $Pr(\phi|\mathbf{s}^n)$ y $Pr(\beta|\mathbf{s}^n)$ son las probabilidades a priori del primer plano y del fondo.

Probabilidades a priori En este caso, a diferencia de la estrategia planteada en el capítulo 5, en la que los modelados del primer plano y del fondo se realizaban teniendo en cuenta información espacial y de apariencia de los píxeles, el fondo ha sido modelado teniendo en cuenta únicamente la información de apariencia. Por ese motivo, las probabilidades a priori utilizadas en el clasificador se calculan como:

$$Pr(\phi|\mathbf{s}^n) = \frac{Pr_\phi(\mathbf{s}^n)}{Pr_\phi(\mathbf{s}^n) + Pr_\beta(\mathbf{s}^n)} \tag{6.7}$$

$$Pr(\beta|\mathbf{s}^n) = \frac{Pr_\beta(\mathbf{s}^n)}{Pr_\phi(\mathbf{s}^n) + Pr_\beta(\mathbf{s}^n)} \tag{6.8}$$

donde $(Pr_\phi(\mathbf{s}^n), Pr_\beta(\mathbf{s}^n))$ son las probabilidades obtenidas del modo descrito en la sección 5.5.2 del capítulo 5, a partir de la predicción resultante de la aplicación del filtro de partículas.

Aplicando estas probabilidades a priori, la probabilidad de que el píxel p^n pertenezca a un objeto móvil se obtiene como:

$$Pr(\phi|\mathbf{x}^n) = \frac{Pr_\phi(\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n,\phi)}{Pr_\phi(\mathbf{s}^n)p(\mathbf{c}^n|\mathbf{s}^n,\phi) + Pr_\beta(\mathbf{s}^n)p(\mathbf{c}^n|\beta)} \tag{6.9}$$

6.4. Análisis de regiones de interés

Normalmente, en una secuencia de vídeo, la mayor parte de los píxeles de las imágenes forman parte del fondo. Por lo tanto, si se pudieran identificar las regiones de las imágenes en las que es más probable la aparición de objetos móviles y el proceso de detección de dichos objetos se limitara a esas regiones, se podría obtener un importante ahorro computacional.

Partiendo de esta idea, se ha elaborado una estrategia que permite determinar las regiones que deben ser analizadas en cada instante. Dicha estrategia se basa en combinar un conjunto de regiones dadas por las partículas predichas, resultantes de la aplicación del filtro de partículas utilizado para actualizar las posiciones de las regiones móviles, con regiones obtenidas mediante un innovador y eficiente muestreo aleatorio por ventanas que hemos denominado *Windowed Random Sampling*, *WRS*. Como resultado de esa combinación, para cada imagen I^n, se obtiene una máscara, M^n, de regiones de interés (*Regions of Interest*, *RoI*) que determina los píxeles de la imagen que deben ser procesados.

Por un lado, considerando las regiones cubiertas por las partículas predichas, se tienen en cuenta las localizaciones en las que es más probable la presencia de los objetos móviles que están siendo seguidos con el filtro de partículas. Por otro lado, el *WRS* permite la detección de nuevas regiones móviles (aquellas que todavía no estén siendo contempladas por el filtro de partículas).

Las muestras resultantes del *WRS* son pequeñas regiones dispersas de forma aleatoria en ventanas uniformemente distribuidas en las imágenes. De ese modo, estas regiones cubren un pequeño porcentaje del total de la imágenes y son capaces de detectar la presencia de nuevos objetos móviles en cualquier zona.

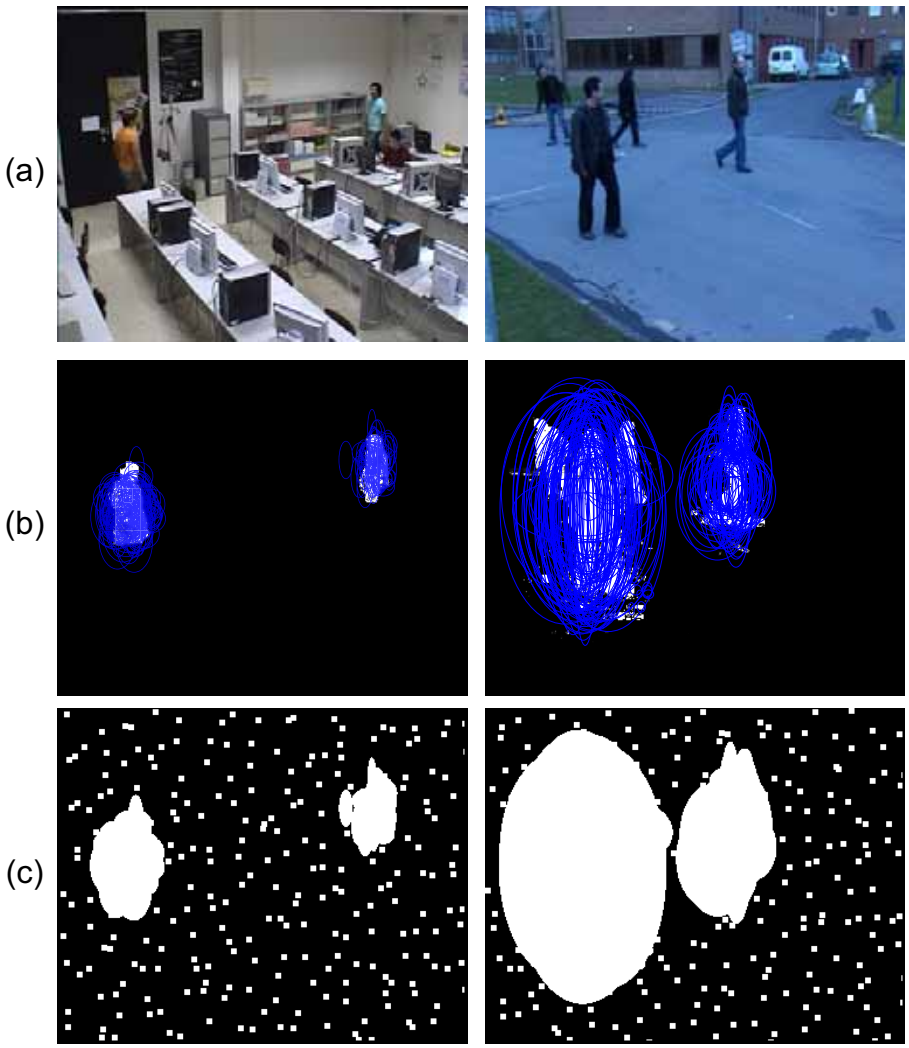


Figura 6.3: Ejemplos de máscaras de regiones de interés. (a) Imágenes originales. (b) Partículas predichas sobre las detecciones obtenidas. (c) Máscaras de regiones de interés.

En la figura 6.3 se han representado dos ejemplos de generación de las máscaras de regiones de interés. La primera fila de imágenes (figura 6.3.a) muestra las imágenes originales analizadas. La segunda fila (figura 6.3.b) presenta las partículas predichas, resultantes de la aplicación del filtro de partículas sobre las últimas detecciones obtenidas. Por último, en la tercera fila de imágenes (figura 6.3.c) se pueden observar las máscaras generadas, las cuales consideran únicamente los píxeles cubiertos por las partículas predichas y por las regiones aleatorias resultantes del *WRS*.

En el momento en el que aparece un nuevo objeto móvil en la escena, inicialmente es detectado como una o varias regiones gracias a las muestras del *WRS*. Estas regiones comenzarán a ser tenidas en cuenta por el filtro de partículas y, consecuentemente sus

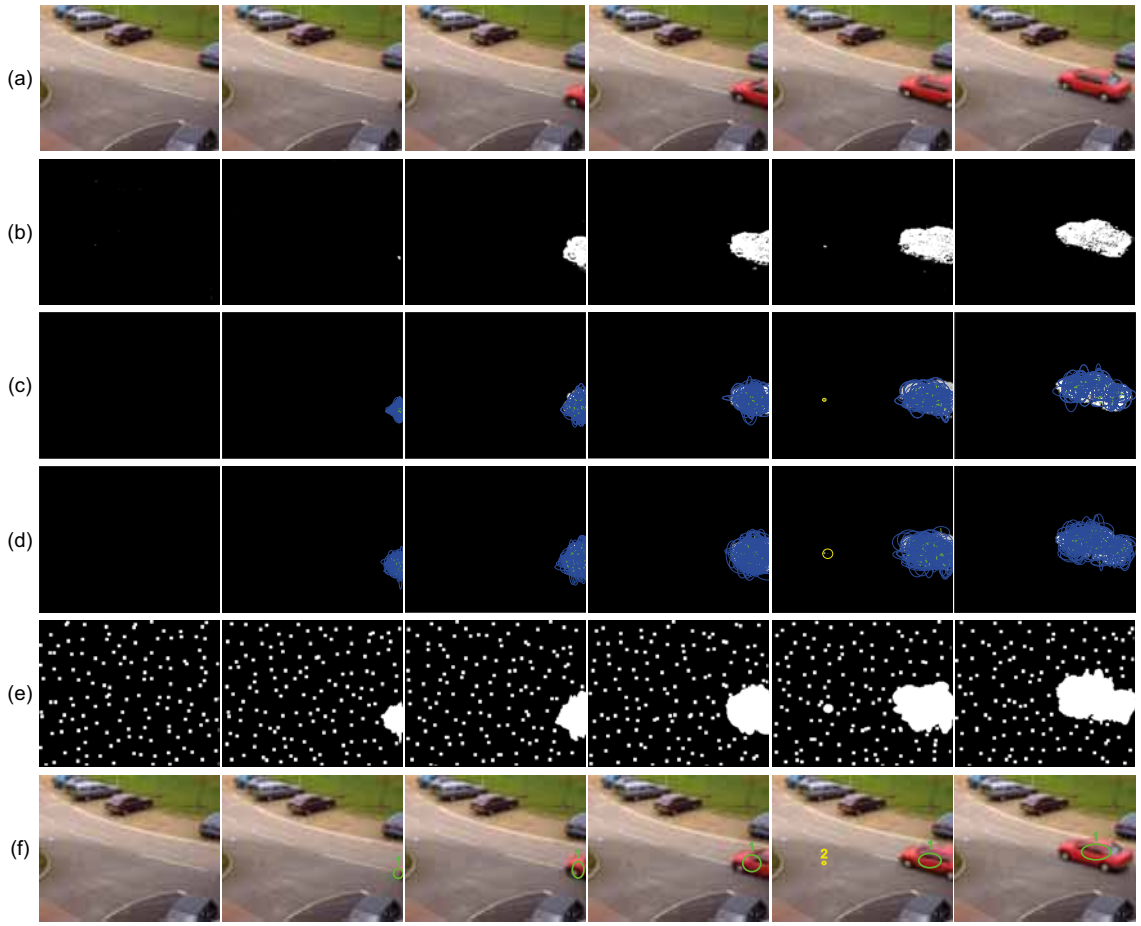


Figura 6.4: Ejemplo de generación de las máscaras de regiones de interés cuando se detecta un nuevo objeto. (a) Imágenes originales. (b) Detecciones. (c) Partículas asociadas a las regiones móviles detectadas. (d) Partículas predichas. (e) Máscaras de regiones de interés. (f) Estimación de los vectores de estado sobre las imágenes originales.

partículas predichas asociadas generarán regiones de interés en su entorno, permitiéndolas crecer hasta que cubran por completo al nuevo objeto móvil.

En la figura 6.4 se ha representado un ejemplo de esta situación. En la primera fila de imágenes (figura 6.4.a) se pueden ver algunas de las imágenes originales de una secuencia en la que se observa que un coche aparece en escena. La segunda fila (figura 6.4.b) presenta las detecciones obtenidas para cada una de estas imágenes. El la tercera fila (figura 6.4.c) se han representado las partículas asignadas a las regiones móviles detectadas. En la cuarta (figura 6.4.d) se muestran las partículas predichas, de las cuales se extrae parte de las regiones de interés. En la quinta fila de imágenes se pueden ver las máscaras de regiones de interés, resultantes de la combinación del *WRS* y de las partículas predichas. Por último, en la última fila se han representado las estimaciones puntuales de los vectores de estado resultantes de la aplicación del filtro de partículas. Estos resultados permiten apreciar cómo

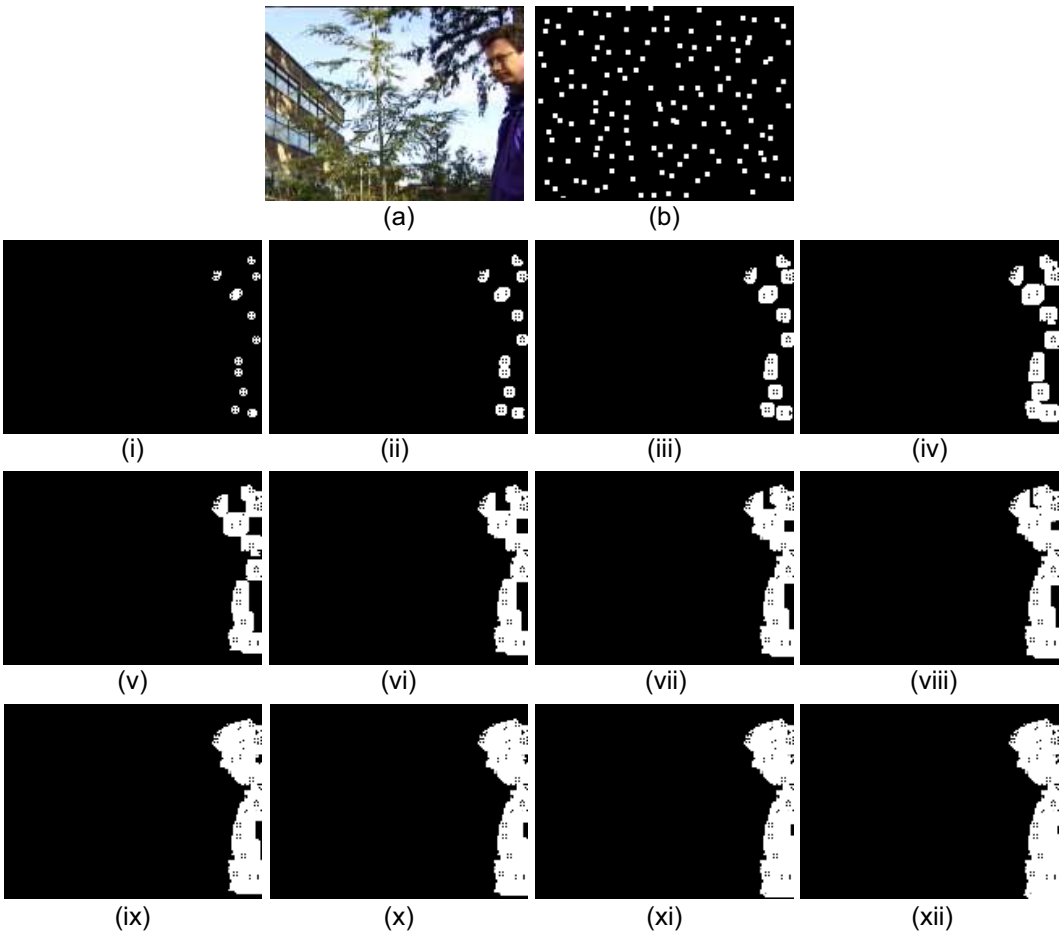


Figura 6.5: Crecimiento de regiones móviles nuevas detectadas con el *WRS*. (a) Imagen original. (b) Máscara de regiones de interés. (i-xii) Resultados de cada iteración a lo largo de la etapa de crecimiento de las regiones inicialmente detectadas.

un nuevo objeto móvil que inicialmente es detectado por una de las muestras aleatorias del *WRS*, tras el análisis de unas cuantas imágenes, acaba dando lugar a un área dentro de la máscara de regiones de interés que resulta lo suficientemente grande como para cubrir dicho objeto por completo. En la quinta columna de imágenes de este ejemplo también se puede observar que aparece un segundo grupo de partículas (representado en color amarillo), resultante de la detección de un segundo objeto móvil gracias a una de las muestras del *WRS*. Sin embargo, al tratarse de una falsa detección, en la siguiente imagen (sexta columna de imágenes) dicho objeto ha desaparecido y, consecuentemente, la región de interés que se había creado para cubrir su posible localización también desaparece.

Cuando un nuevo objeto móvil es detectado por primera vez gracias a las regiones aleatorias resultantes del *WRS*, su área máxima detectada será, cómo máximo, la cubierta por las regiones aleatorias que han permitido tenerlo en cuenta. En estos casos, para conseguir

partir de una detección que se adapte mejor al objeto móvil por completo, se aplica un algoritmo de crecimiento sobre las nuevas regiones móviles detectadas. Este crecimiento se limita a realizar el proceso de detección, de forma iterativa, sobre los píxeles contiguos a aquellos previamente clasificados como parte del primer plano, finalizando cuando ninguno de los nuevos píxeles analizados se a etiquetado como móvil. En la figura 6.5 se ha representado un ejemplo de esta situación. En este ejemplo se ha analizado el instante de aparición de un objeto móvil de grandes dimensiones (figura 6.5.a). En dicho instante, el análisis de las regiones de interés (figura 6.5.b) ha dado lugar a detección inicial de pequeñas porciones del objeto móvil. Sin embargo, después de aplicar el algoritmo de crecimiento sobre estas detecciones iniciales (los resultados correspondientes a cada iteración del algoritmo se han representado en el resto de imágenes de la figura) la detección obtenida considera la mayor parte del objeto móvil.

6.5. Resultados

Para evaluar las mejoras obtenidas con las estrategias propuestas en el presente capítulo se ha llevado a cabo un análisis de su eficiencia computacional y de la calidad de los resultados que proporciona. Dichos análisis se han efectuado sobre el conjunto de secuencias descrito en la sección A.2 del apéndice A. Al igual que en el capítulo 5, para modelar el fondo de las secuencias se han utilizado $N_{\beta_N} = 150$ imágenes de referencia, mientras que para modelar el primer plano se han utilizado únicamente $N_{\phi_N} = 10$ imágenes de referencia. Además, para reducir la presencia de falsas detecciones debidas a pequeños movimientos de la cámara o a píxeles del fondo con variaciones ruidosas, se ha aplicado una apertura morfológica sobre las detecciones (Gao et al., 2008).

En primer lugar, en la sección 6.5.1 se ofrece un análisis detallado de la mejora computacional lograda con la estrategia propuesta. Acto seguido, en la sección 6.5.2 se analiza la calidad de los resultados obtenidos y se compara con los resultados obtenidos mediante las estrategias propuestas en los capítulos 4 y 5.

6.5.1. Análisis computacional

Las estrategias de detección que utilizan la información espacial de los píxeles de referencia para construir los modelos correspondientes al fondo y al primer plano (por ejemplo, la propuesta en el capítulo 5, o las propuestas en trabajos como (Sheikh y Shah, 2005) y (Zhang y Yang, 2008)) tienen asociado un elevado coste computacional. Este coste es especialmente significativo en el caso del modelado del fondo, ya que se lleva a cabo a partir de una gran cantidad de muestras de referencia que, para cada píxel de la imagen actual, han de ser evaluadas mediante *kernels* multidimensionales.

Sin embargo, en este capítulo se plantea la posibilidad de aplicar una estrategia de detección en la que la información espacial se utiliza únicamente en el modelado del primer plano. De esta forma, para modelar cada uno los píxeles de la imagen actual, sólo es necesario evaluar tantos *kernels* como imágenes de referencia se estén utilizando.

En la tabla 6.1 se han representado las cantidades típicas de *kernels* que han de evaluarse

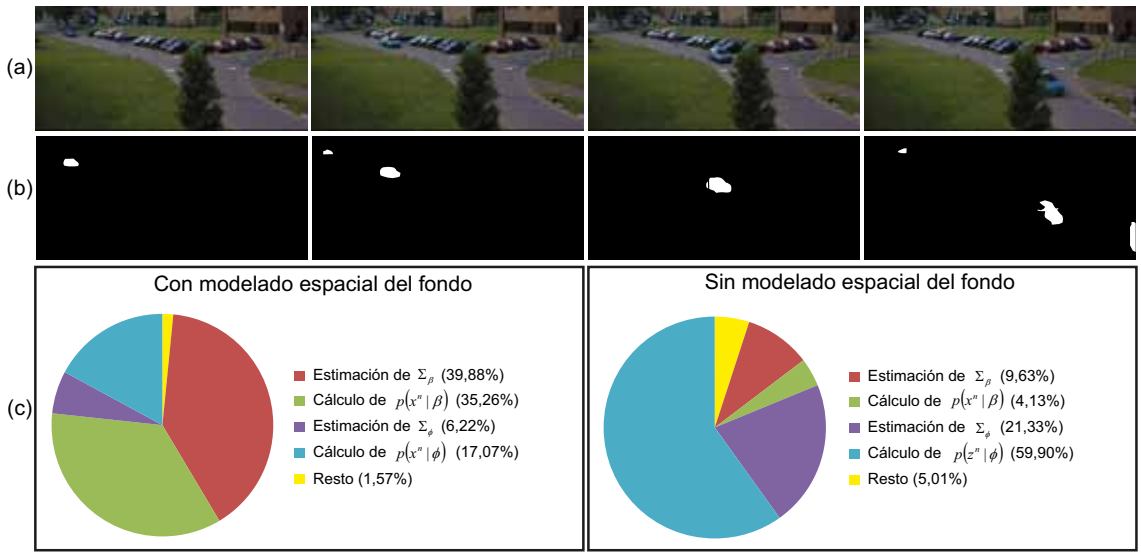


Figura 6.6: Porcentaje de coste computacional asociado a cada etapa de la estrategia de detección propuesta. (a) Imágenes representativas de la secuencia analizada. (b) Detecciones de referencia. (c) Porcentajes obtenidos con la estrategia propuesta en el capítulo 5 (gráfico de la izquierda) y con la estrategia descrita en el presente capítulo (gráfico de la derecha).

	Coste	Ratio
Estrategias tradicionales	$\propto N_\beta HW$	10^5
Nuestra estrategia	$\propto N_\beta$	1

Tabla 6.1: Coste computacional, a nivel de píxel, correspondiente al modelado del fondo.

para llevar a cabo el modelado del fondo de un píxel en una secuencia constituida por imágenes de $H \times W = 280 \times 352$ píxeles. Los resultados de esta tabla muestran que con la estrategia descrita en este capítulo, al no utilizarse información espacial en el modelado, el número de operaciones requeridas es 5 órdenes de magnitud inferior al utilizado en las estrategias de modelado no paramétrico tradicionales (Sheikh y Shah, 2005) y (Zhang y Yang, 2008).

En la figura 6.6 se puede observar un análisis del coste computacional asociado a cada una de las etapas de las que constan las estrategias propuestas en el capítulo 5 y en el presente capítulo, tras su aplicación sobre una secuencia de 435 imágenes. Algunas de las imágenes características de dicha secuencia y sus detecciones de referencia correspondientes se han representado, respectivamente, en figura 6.6.a y en la figura 6.6.b. Por un lado, en el gráfico de izquierda de la figura 6.6.c se ha representado el porcentaje de tiempo utilizado en las distintas etapas de la estrategia propuesta en el capitulo 5. Por otro lado, en el gráfico de la derecha de la figura 6.6.c se muestran los porcentajes correspondientes a esas mismas etapas, con las modificaciones propuestas en este capítulo.

Observando los datos del gráfico de la izquierda se puede comprobar que aproxima-

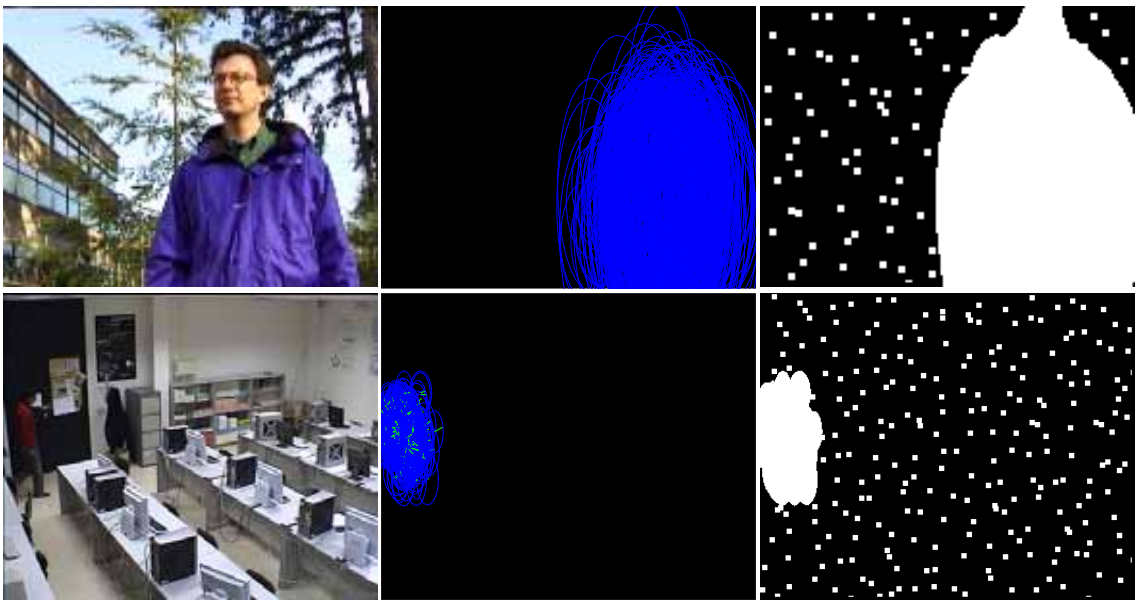


Figura 6.7: Regiones de interés en secuencias con distinta proporción de píxeles móviles. (a) Imágenes originales. (b) Partículas resultantes de la aplicación del filtro de partículas. (c) Máscaras de regiones de interés.

damente el 75 % del coste computacional se debe a las etapas asociadas al modelado del fondo (un 39,88 % en la estimación de las matrices de escala de los *kernels* y un 35,6 % en la evaluación de los *kernels*). El 25 % restante se corresponde con la estimación de las matrices de escala del primer plano (6,22 %), el modelado del primer plano (17,07 %) y el resto de módulos del sistema (obtención de la información a priori, clasificador bayesiano y filtro de partículas), los cuales únicamente suponen el 1,57 % del coste total. Hay que tener en cuenta que estos resultados se han obtenido aplicando los márgenes espaciales definidos en la sección 5.4.1, los cuales delimitan la cantidad de información espacial utilizada en el modelado del fondo, ya que si se aplicara directamente la estrategia de modelado propuesta en (Sheikh y Shah, 2005) o en (Zhang y Yang, 2008) el coste computacional de esta etapa sería muy superior (tal y como se muestra en la tabla 6.1, unos 5 órdenes de magnitud).

Prestando ahora atención a los datos correspondientes a la estrategia descrita en este capítulo (mostrados en el gráfico de la derecha de la figura 6.6.c), se observa que las fases que inicialmente eran las más costosas (estimación de las matrices de escala de los *kernels* y cálculo de la densidad de probabilidad del fondo) ahora sólo representan el 14 % del coste computacional total. Por lo tanto, aplicando la estrategia propuesta sobre la secuencia analizada en este ejemplo, el coste computacional total se ha reducido aproximadamente un 70 % (suponiendo que el coste asociado al resto de etapas se ha mantenido aproximadamente constante).

Por otro lado, además del ahorro computacional obtenido al no utilizar información espacial para modelar fondo, gracias a la aplicación de las máscaras de regiones de interés

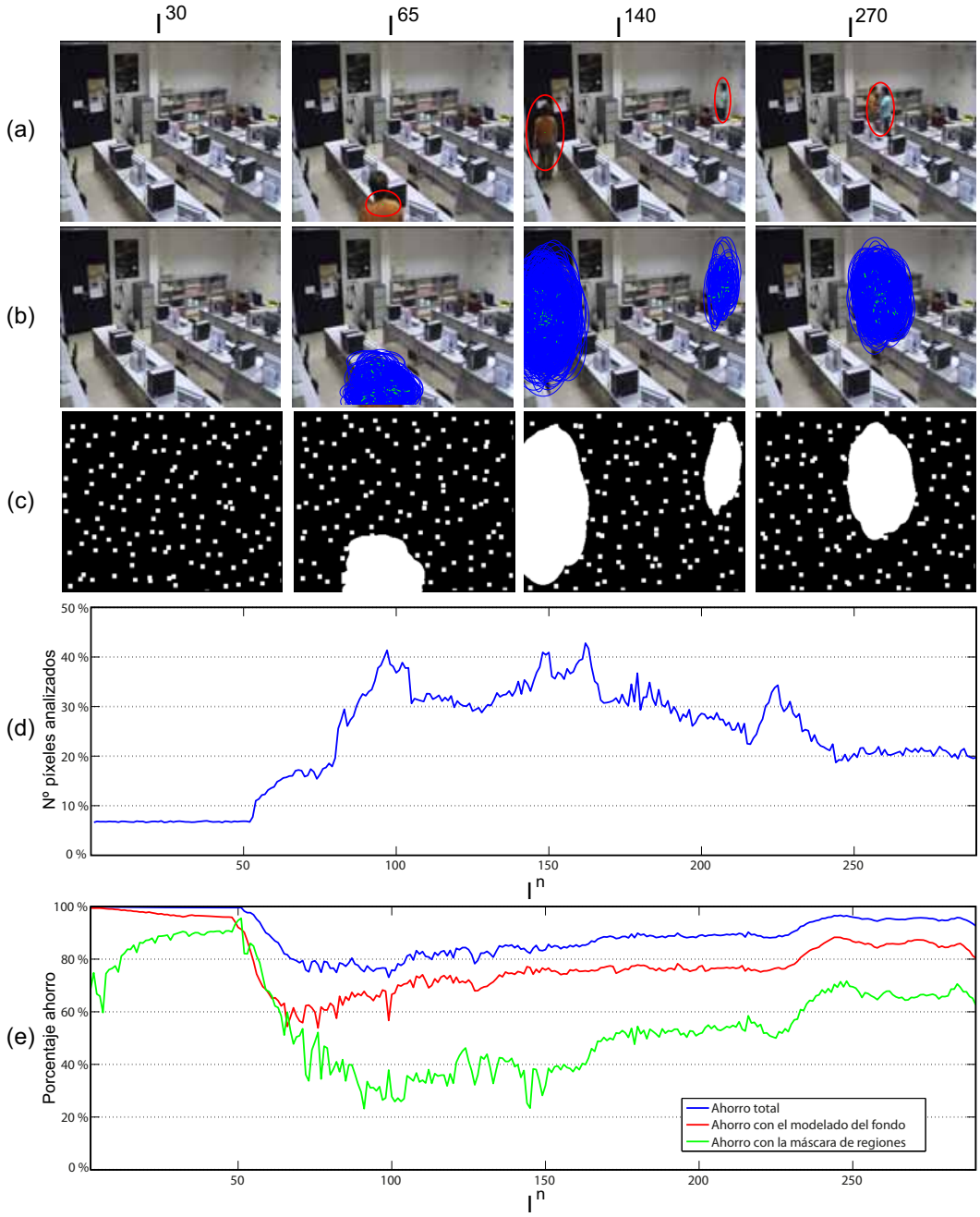


Figura 6.8: Ahorro computacional obtenido sobre una secuencia de 290 imágenes. (a) Estimación puntual de los vectores de estado sobre algunas imágenes de la secuencia. (b) Partículas predichas, resultantes de la aplicación del filtro de partículas. (c) Máscaras de regiones de interés. (d) Porcentaje de píxeles analizados en cada imagen. (e) Porcentaje de ahorro computacional en cada imagen.

	Píxeles móviles (%)	Píxeles analizados (%)	Ahorro 1 (%)	Ahorro 2 (%)
Lab_001	1,00	14,08	77,60	88,79
Lab_002	2,47	23,61	71,90	85,26
Lab_003	2,72	15,86	82,98	90,84
Lab_004	2,79	17,73	69,32	75,89
Lab_005	0,94	15,44	81,94	92,66
Lab_006	5,25	33,94	65,81	71,53
Pets_001	1,12	19,78	88,77	96,74
Pets_002	0	7,76	93,57	99,47
Pets_003	0,37	11,44	87,48	96,81
Pets_004	5,40	68,22	68,98	73,93
Pets_005	6,27	60,43	62,80	70,44
Wall_001	6,79	17,89	33,40	37,06
Wall_002	1,52	9,39	79,78	84,01

Tabla 6.2: Ahorro computacional obtenido mediante la utilización de las máscaras de regiones de interés.

se consigue un ahorro aún mayor. Este ahorro, como ya se ha mencionado anteriormente (sección 6.4), es proporcional a la cantidad de píxeles móviles previamente identificados como tales y, por lo tanto, depende del contenido móvil de cada secuencia. Las secuencias con objetos móviles grandes darán lugar grandes regiones de interés (ejemplo mostrado en la primera fila de imágenes de la figura 6.7) y, por lo tanto, requerirán analizar mayor porcentaje de píxeles que las secuencias en las que, por el menor tamaño de sus objetos móviles, las regiones de interés sean más pequeñas (ejemplo mostrado en la segunda fila de imágenes de la figura 6.7). Además, también se ha de tener en cuenta que el ahorro computacional obtenido mediante la utilización de las máscaras de regiones de interés normalmente no es constante, ya que la cantidad de píxeles móviles a lo largo de una secuencia tampoco lo es.

En la figura 6.8 se muestra el ahorro computacional obtenido a lo largo de una secuencia de 290 imágenes. En la primera fila (figura 6.8.a) aparecen algunas imágenes representativas de la secuencia analizada. En la segunda fila (figura 6.8.b) se pueden ver las detecciones de referencia de dichas imágenes y en la tercera fila (figura 6.8.c) se muestran sus máscaras de regiones de interés. En la cuarta fila (figura 6.8.d) se ha representado el porcentaje de píxeles analizados (aquellos indicados por las máscaras de regiones de interés) a lo largo de toda la secuencia. En la última fila (figura 6.8.e) se muestran diferentes porcentajes de ahorro computacional: el obtenido al eliminar la información espacial en el modelado del fondo (gráfica de color rojo), el resultante de la aplicación de las máscaras de regiones de interés (gráfica de color verde) y, por último, el que resulta de la aplicación conjunta de ambas estrategias de ahorro (gráfica de color azul). Analizando los datos de estas gráficas (figura 6.8.d y figura 6.8.e) se puede apreciar que el ahorro computacional obtenido depende de la cantidad de píxeles móviles presentes en la escena, siendo menor a medida que esta cantidad de píxeles es mayor: en el caso del ahorro obtenido por haber modelado el fondo

	Estrategia propuesta en el capítulo 4			**Estrategia propuesta en el capítulo 5**		
	Recall	Precision	**F**	Recall	Precision	**F**
Lab_001	73,19	83,92	78,19	96,88	72,06	82,64
Lab_002	76,39	84,37	80,18	97,68	60,08	74,40
Lab_003	77,56	65,72	71,15	71,37	63,73	67,34
Lab_004	83,60	69,24	75,75	95,96	66,35	78,46
Lab_005	69,03	84,09	75,82	93,59	63,18	75,43
Lab_006	91,20	58,98	71,63	78,21	60,09	67,96
Pets_001	87,54	71,77	78,87	87,93	26,12	40,27
Pets_002	100	0	0	100	0	0
Pets_003	81,49	69,21	74,85	91,24	76,44	83,19
Pets_004	74,37	81,94	77,97	85,09	51,29	64,01
Pets_005	82,34	90,28	86,13	94,97	53,32	68,30
Wall_001	95,49	43,83	60,09	95,90	79,61	87,00
Wall_002	95,55	6,72	12,56	93,71	77,05	84,57
Promedio	86,78	20,84	33,60	91,72	59,16	71,93

	Estrategia propuesta en el presente capítulo (sin máscaras de RoI)			**Estrategia propuesta en el presente capítulo (con máscaras de RoI)**		
	Recall	Precision	**F**	Recall	Precision	**F**
Lab_001	95,20	88,36	91,66	95,22	88,14	91,55
Lab_002	93,35	86,85	89,98	92,78	86,80	89,69
Lab_003	74,64	86,92	80,32	74,60	87,48	80,53
Lab_004	95,91	92,60	94,23	95,91	92,73	94,29
Lab_005	94,42	82,36	87,98	91,83	82,39	86,85
Lab_006	75,47	72,48	73,95	77,05	71,07	73,94
Pets_001	85,08	81,08	83,03	84,20	81,57	82,87
Pets_002	100	0	0	100	0	0
Pets_003	91,61	83,10	87,15	88,76	83,33	85,96
Pets_004	82,24	91,05	86,42	81,35	91,21	85,99
Pets_005	91,71	88,14	89,89	91,41	89,42	90,40
Wall_001	97,90	97,98	97,94	97,38	98,25	97,81
Wall_002	95,34	94,11	94,72	93,96	98,30	96,08
Promedio	91,11	91,13	91,12	90,62	91,69	91,15

Tabla 6.3: Resumen de la calidad obtenida con la estrategia propuesta es este capítulo, comparada con la obtenida mediante las estrategias descritas en los capítulos 4 y 5.

sin considerar la información espacial de los píxeles, esto se debe a que cuanta más cantidad de píxeles móviles hay en escena, menor es el porcentaje de coste computacional asociado al modelado del fondo y, por lo tanto, menos apreciable es la mejora computacional obtenida

en dicho modelado; mientras que en el caso del ahorro obtenido con las máscaras de regiones de interés se debe a que el área cubierta por las regiones de interés es proporcional a la cantidad de píxeles móviles.

Para terminar, en la tabla 6.2 se ha resumido el ahorro computacional logrado en el análisis de todas las secuencias de la base de datos utilizada (en relación con la estrategia de detección propuesta en el capítulo 5). La primera columna de dicha tabla contiene los nombres de las secuencias. La segunda columna muestra el porcentaje medio de píxeles móviles en cada secuencia. La tercera columna muestra el porcentaje medio de píxeles analizados (determinado por el área cubierta por las regiones de interés). La penúltima columna presenta el ahorro computacional obtenido sin hacer uso de las máscaras de regiones de interés (debido únicamente a la estrategia de ahorro computacional en el modelado del fondo), mientras que la última columna muestra el ahorro obtenido tras añadir la etapa de estimación de las máscaras de regiones de interés. Como se puede ver en estos resultados, el ahorro computacional es mayor en las secuencias con menor proporción de contenido móvil, obteniéndose ahorros de hasta el $99,47\,\%$ en secuencias como $Pets\_002$, en la que el contenido móvil es nulo. Sin embargo, en secuencias en las que el porcentaje de píxeles móviles es notablemente superior (por ejemplo, $Wall\_001$), debido a que la cantidad de píxeles descartados por las máscaras es muy pequeña y a que el porcentaje de coste computacional asociado al modelado del primer plano es muy superior, se obtiene menor ahorro computacional.

6.5.2. Análisis de calidad

Al igual que se ha hecho en los capítulos previos, para analizar la calidad de la estrategia de detección descrita en este capítulo se ha hecho uso de los porcentajes de *Recall*, *Precision* y F descritos en la sección 3.7 del capítulo 3, calculados del modo detallado en la sección 4.4 del capítulo 4. En la tabla 6.3 se puede ver el resumen de dicho análisis en dos situaciones: haciendo uso de las máscaras de regiones de interés y sin utilizar dichas máscaras. Además, para poder analizar la mejora obtenida frente a las estrategias previamente descritas a lo largo de esta tesis, dicha tabla también muestra el resumen de resultados correspondientes a las estrategias de detección propuestas en los capítulos 4 y 5.

Analizando los resultados mostrados en dicha tabla se puede comprobar que con la estrategia propuesta, gracias a la utilización de las componentes de color normalizadas junto con el módulo del gradiente de la saturación, se ha conseguido una importante reducción del número de falsas detecciones (aumento de los valores de *Precision*) debida, principalmente, a la supresión de las sombras y los reflejos en las detecciones. En las figuras 6.9 y 6.10 se han representado algunos de los resultados obtenidos en situaciones en las que los objetos móviles generan sombras y reflejos en sus desplazamientos: en escenarios cerrados, en el caso de la figura 6.9; y en escenarios al aire libre, en el de la figura 6.10. La primera fila de imágenes de ambas figuras (a) muestra las imágenes originales analizadas. La segunda fila (b) contiene las detecciones de referencia de dichas imágenes. En la tercera fila (c) se han representado los resultados obtenidos con el método de mezcla de gaussianas propuesto en el capítulo 4. La cuarta fila (d) muestra las detecciones obtenidas con el método no paramétrico propuesto en el capítulo 5 y las dos últimas filas contienen las detecciones

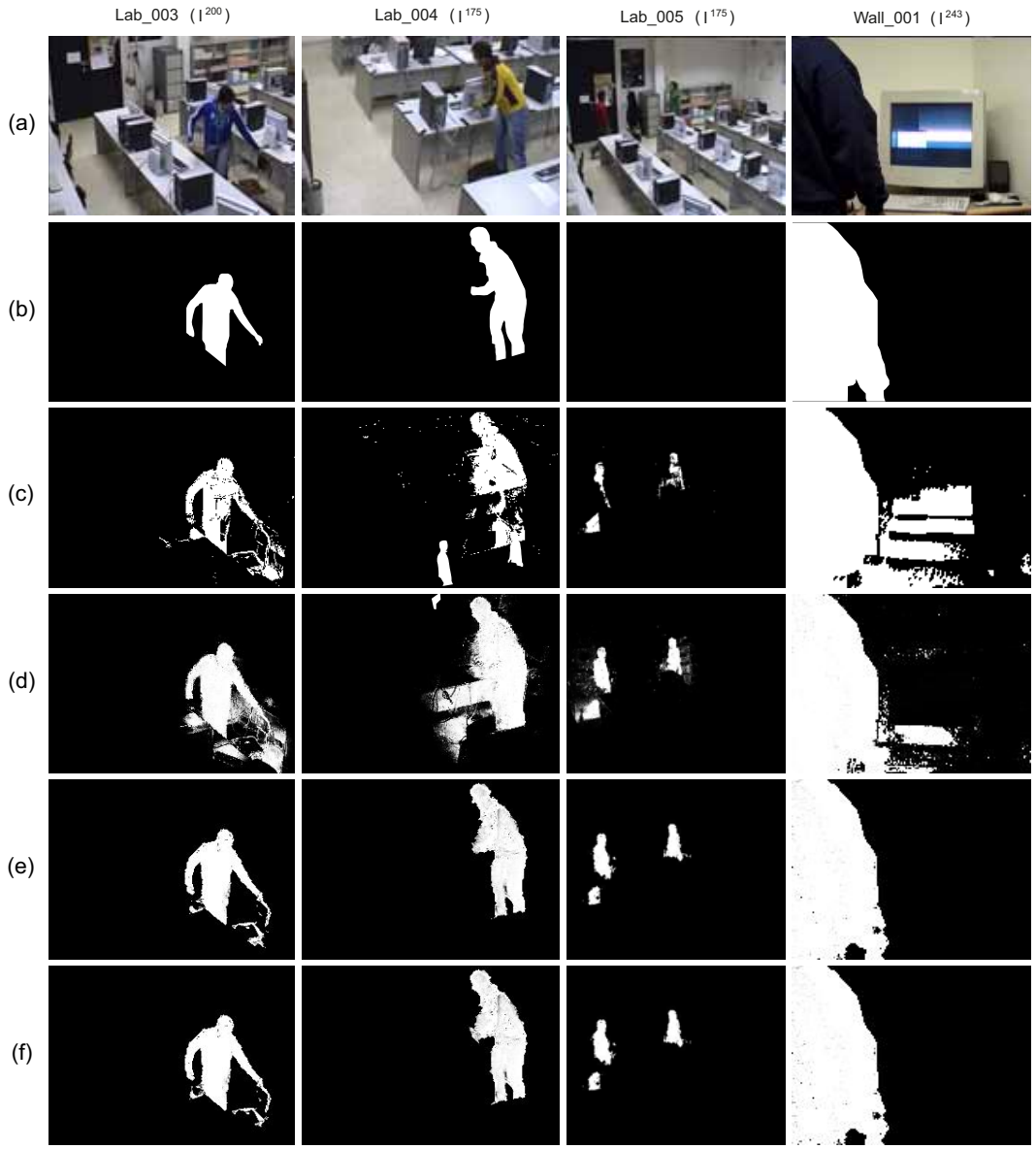


Figura 6.9: Detecciones obtenidas en escenarios cerrados. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con la estrategia propuesta en el capítulo 4. (d) Detecciones obtenidas con la estrategia propuesta en el capítulo 5. (e) Detecciones obtenidas con la estrategia propuesta en el presente capítulo, sin hacer uso de las máscaras de regiones de interés. (f) Detecciones obtenidas con la estrategia propuesta en el presente capítulo, utilizando las máscaras de regiones de interés.

resultantes de la aplicación de la estrategia descrita en el presente capítulo: sin utilizar las máscaras de regiones de interés (e) y haciendo uso de dichas máscaras (f). Comparando los resultados mostrados en estas dos figuras es fácil apreciar que con la estrategia descrita en

Figura 6.10: Detecciones obtenidas en escenarios al aire libre. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con la estrategia propuesta en el capítulo 4. (d) Detecciones obtenidas con la estrategia propuesta en el capítulo 5. (e) Detecciones obtenidas con la estrategia propuesta en el presente capítulo, si hacer uso de las máscaras de regiones de interés. (f) Detecciones obtenidas con la estrategia propuesta en el presente capítulo, utilizando las máscaras de regiones de interés.

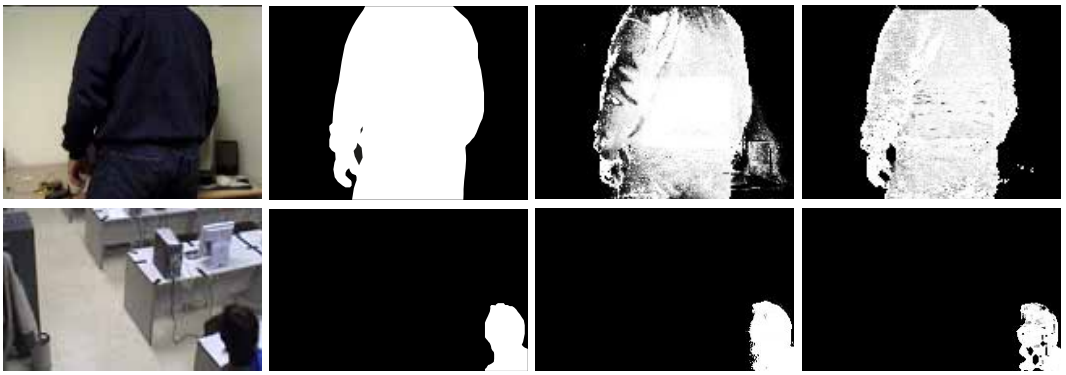


Figura 6.11: Análisis de la calidad de las detecciones en función de las características de apariencia utilizadas en los modelados. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas con las componentes *RGB* utilizadas en la estrategia propuesta en el capítulo 5. (d) Detecciones obtenidas con el conjunto de componentes propuesto en la estrategia descrita en el presente capítulo.

este capítulo, tanto si se aplican las máscaras de regiones de interés como si no se aplican, se obtiene una reducción muy significativa del número de falsas detecciones debidas a las sombras y a los reflejos.

Volviendo a los resultados mostrados en la tabla 6.3, si se presta ahora atención a los porcentajes de *Recall* obtenidos con la estrategia propuesta en este capítulo y se comparan con los obtenidos mediante la estrategia propuesta en el capítulo 5, se puede apreciar que en algunos casos han aumentado ligeramente y en otros casos han disminuido también ligeramente. Los casos para los que han aumentado son aquellos en los que el conjunto de componentes de apariencia propuestas para llevar a cabo los modelados (color normalizado y módulo del gradiente de la saturación) permite discriminar mejor entre fondo y primer plano que las componentes *RGB*. En la primera fila de imágenes de la figura 6.11 se ha representado un ejemplo de esta situación en la que se comprueba que el objeto móvil presente en la escena se detecta mejor con el conjunto de componentes propuesto en este capítulo. Por otro lado, los casos en los que el porcentaje de *Recall* se ha reducido se deben, principalmente, al caso opuesto: situaciones en las que las componentes *RGB* aportan más información para discriminar entre fondo y primer plano. Un ejemplo de una situación de este tipo es que se ha representado en la segunda fila de imágenes de la figura 6.11, en la que claramente se puede apreciar que con la combinación de color normalizado y el módulo del gradiente de la saturación se obtiene peor resultado (mayor número de píxeles móviles no detectados) que haciendo uso de las componentes *RGB*.

Si se comparan ahora los resultados obtenidos sin aplicar las máscaras de regiones de interés y aplicándolas, se puede apreciar que cuando se utilizan las máscaras el porcentaje de *Recall* se reduce ligeramente en la mayor parte de las secuencias, mientras que el de *Precision* aumenta también ligeramente en casi todas ellas. La disminución del porcentaje de *Recall* se debe a que al utilizarse las máscaras de regiones de interés, en algunos casos,

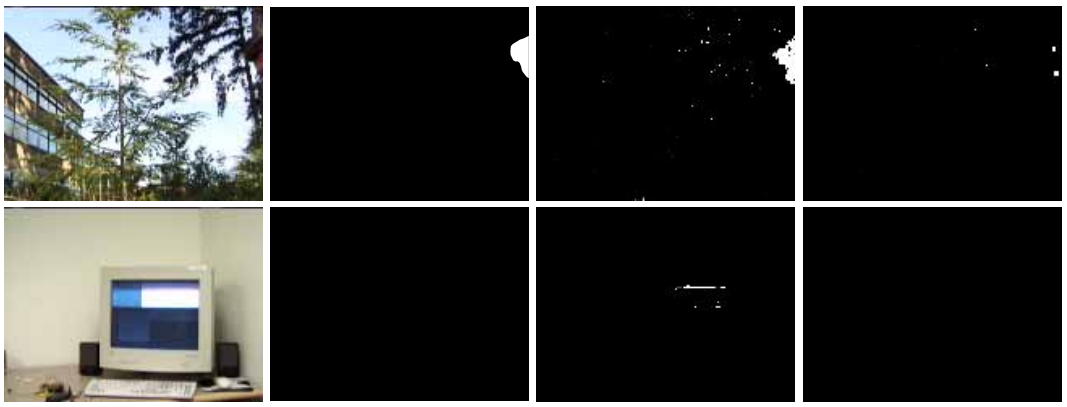


Figura 6.12: Análisis de la influencia de las máscaras de regiones de interés en la calidad de las detecciones. (a) Imágenes originales. (b) Detecciones de referencia. (c) Detecciones obtenidas sin hacer uso de las máscaras de regiones de interés. (d) Detecciones obtenidas con con las máscaras de regiones de interés.

cuando un objeto móvil entra en escena no es inmediatamente detectado, siendo necesarias varias imágenes en las que aparezca dicho objeto para que alguna de las regiones del WRS lo detecte. En la primera fila de imágenes de la figura 6.12 se muestra un ejemplo de esta situación en la que, en el instante correspondiente a la aparición de un objeto móvil, la máscara de regiones de interés no lo ha tenido en cuenta y, consecuentemente, no ha sido detectado. Sin embargo, en dicho ejemplo también se puede apreciar que si no se utiliza la máscara de regiones de interés la detección del objeto móvil es correcta. Por otro lado, al igual que la detección de un nuevo objeto móvil depende de las regiones aleatorias resultantes del WRS, el uso de dichas regiones también reduce la detección de falsos positivos y, por lo tanto, es la responsable de que el porcentaje de *Precision* aumente en la mayor parte de las secuencias. La segunda fila de imágenes de la figura 6.12 muestra un ejemplo que permite comparar los resultados obtenidos, con y sin máscaras de regiones de interés, en una secuencia con un fondo muy dinámico. Se puede observar que haciendo uso de las máscaras se ha evitado la aparición de las falsas detecciones que sí que aparecen cuando no se utilizan.

6.6. Conclusiones

En este capítulo se ha descrito un algoritmo para la detección de objetos móviles, basada en la estrategia de modelado no paramétrico del fondo y del primer plano presentada en el capítulo 5, capaz de: por un lado, reducir muy apreciablemente las falsas detecciones debidas a las sombras y a los reflejos que provocan los objetos móviles en sus desplazamientos; y, por otro lado, mejorar muy significativamente la eficiencia computacional en comparación con la estrategia en la que se basa.

Para modelar tanto el fondo como el primer plano se ha propuesto el uso de un novedoso y efectivo conjunto de características de apariencia de los píxeles: sus componentes de

color normalizadas y el módulo del gradiente de la saturación. Las componentes de color normalizadas consiguen reducir la influencia de los cambios de luminosidad que provocan las sombras y los reflejos, mientras que el módulo del gradiente de la saturación evita perder la información de luminancia de los objetos móviles. De ese modo se ha conseguido obtener detecciones de gran calidad en las que los objetos móviles han sido adecuadamente identificados y en las que el número de falsas detecciones debidas a sombras y reflejos se ha reducido muy significativamente en comparación con los resultados obtenidos mediante otros conjuntos de características de apariencia.

Para aumentar la eficiencia computacional se han aplicado dos estrategias de mejora: una que limita el uso de la información espacial de los píxeles a únicamente el modelado del primer plano; y otra que estima máscaras de regiones de interés que, en cada instante, determinan los píxeles que deben ser analizados. La primera de estas dos estrategias parte del hecho de que el mayor coste computacional se debe a las etapas asociadas al modelado del fondo y, por lo tanto, reduciendo la cantidad de muestras de referencia utilizadas para llevarlo a cabo, hace posible obtener importantes mejoras computacionales. La segunda estrategia de mejora combina pequeñas regiones aleatoriamente distribuidas en ventanas uniformemente repartidas en cada imagen, con las regiones cubiertas por las partículas predichas resultantes de la aplicación del filtro de partículas utilizado para actualizar las posiciones de los objetos móviles. De ese modo, con las primeras se consigue detectar los nuevos objetos móviles que entran en escena y con las segundas se consigue tener en cuenta las regiones en las que es más probable la presencia de los objetos móviles que han sido detectados previamente. Como resultado de la aplicación de estas dos estrategias de mejora se han logrado importantes ahorros computacionales.

Para evaluar la calidad de los resultados obtenidos, así como su mejora computacional, se han analizado numerosas secuencias con contenido especialmente crítico para los sistemas de detección de objetos móviles: fondos dinámicos, sombras y reflejos, objetos móviles parecidos al fondo, etc. Además, todos los resultados obtenidos han sido comparados los los obtenidos mediante las estrategias propuestas en los capítulos previos, comprobándose un importante aumento de la calidad de los resultados (debido principalmente a la supresión de las sombras y los reflejos en las detecciones) y obteniéndose mejoras computacionales muy significativas.

Capítulo 7

Conclusiones y trabajo futuro

El futuro tiene muchos nombres.
Para los débiles es lo inalcanzable.
Para los temerosos, lo desconocido.
Para los valientes es la oportunidad.
Victor Hugo (1802-1885),
novelista francés.

A lo largo de esta tesis se han descrito y analizado distintas estrategias que permiten segmentar secuencias de vídeo en tomas y detectar los objetos móviles presentes en las mismas. En este último capítulo se resumen tanto las principales características de estas estrategias, como sus ventajas frente a otras estrategias con propósitos similares. Además, se plantean algunas opciones de trabajo futuro que permitan evolucionar sobre los distintos esquemas de segmentación y detección propuestos.

7.1. Conclusiones

Como resultado de los importantes avances tecnológicos experimentados durante los últimos años han surgido numerosas aplicaciones orientadas al análisis de secuencias de vídeo. Por un lado han aparecido herramientas de edición de vídeo que, para poder llevar a cabo la indexación y la búsqueda de contenidos, requieren de estrategias capaces de detectar eficientemente las transiciones que separan las tomas de las que se componen los vídeos. Por otro lado, los numerosos dispositivos de última generación que incluyen cámaras de vídeo precisan de aplicaciones en las que la detección de objetos móviles es una etapa clave para la realización de tareas de más alto nivel. Como respuesta a esta demanda, en esta tesis se han propuesto distintas estrategias que, mejorando el actual estado del arte, cumplen los requisitos de calidad, velocidad y facilidad de uso requeridos por los usuarios de todas estas aplicaciones. A continuación se resumen brevemente las distintas propuestas realizadas.

- **Una estrategia para segmentar las secuencias de vídeo en tomas:** Dicha estrategia, descrita en el capítulo 3, permite detectar muy eficientemente las transiciones, tanto abruptas como graduales, que sirven de frontera entre las tomas que constituyen

las secuencias de vídeo. A diferencia de otras estrategias de segmentación de secuencias, la estrategia propuesta es capaz de proporcionar resultados de gran calidad (detectando la mayor parte de las transiciones y evitando las falsas detecciones), a la vez que mantiene los requisitos computacionales demandados. En una primera etapa se aplican técnicas muy rápidas que permiten detectar prácticamente todas las transiciones (porcentajes de *Recall* superiores al 99 % en el caso de las transiciones abruptas y por encima del 93 % en el caso de las graduales). Acto seguido, en una segunda etapa se analiza el movimiento entre los pares de imágenes que delimitan las transiciones obtenidas en la primera etapa. Este análisis, al realizarse exclusivamente sobre algunos pares de imágenes preseleccionadas, apenas incrementa el coste computacional del sistema y, sin embargo, permite detectar y descartar la mayor parte de las falsas transiciones (proporcionando valores de *Precision* superiores al 96 %). El resultado es un sistema capaz de generar muy buenos resultados y que, además, es capaz de trabajar a gran velocidad.

- **Una estrategia de detección de objetos móviles con mezclas de gaussianas:** Aparece descrita en el capítulo 4 y está basada en el popular método de mezcla de gaussianas (frecuentemente utilizado, gracias a su capacidad para proporcionar buenos resultados en gran cantidad de escenarios y en presencia de fondos que no permanecen estáticos). La principal aportación de dicha estrategia es su capacidad para adaptar dinámicamente el número de gaussianas utilizadas por cada píxel en cada instante. De ese modo permite reducir muy notablemente el número de operaciones que se deben realizar sobre cada imagen y, así, es capaz de obtener grandes mejoras computacionales. Además, gracias a la utilización de un contador que indica cuánto tiempo lleva sin ser utilizada cada gaussiana, el algoritmo propuesto es capaz de reducir la dependencia de los resultados con los valores asignados a algunos de los parámetros que se utilizan, evitando la necesidad de buscar valores adecuados para esos parámetros en función de las características de cada secuencia y, por lo tanto, reduciendo la complejidad de uso del método original.
- **Un sistema de detección de objetos móviles mediante técnicas de modelado no paramétrico:** Aunque la estrategia propuesta en el capítulo 4 es realmente rápida y proporciona resultados de calidad en un gran número de escenarios, presenta algunas limitaciones que se deben tener en cuenta: no es capaz de modelar adecuadamente las variaciones de fondos muy dinámicos, depende de numerosos parámetros y, además, en situaciones en las que el fondo y el primer plano son parecidos, la calidad de las detecciones es insuficiente. Es por eso que en el capítulo 5, como alternativa a esta estrategia, se ha propuesto otro sistema de detección que se basa en técnicas de modelado no paramétrico y que incluye varias aportaciones:
  - **Modelado del fondo y del primer plano utilizando información de apariencia y de posición:** A diferencia de otros métodos de detección, para reducir la cantidad de píxeles móviles no detectados en situaciones en las que el fondo y el primer plano son parecidos, esta estrategia modela tanto el fondo como el primer plano de las secuencias y, además, no sólo utiliza información de apariencia de los píxeles, sino que también hace uso de información relativa a su posición dentro de las imágenes.

- **Estimación dinámica del ancho de los *kernels*:** Adicionalmente, se han propuesto dos novedosas estrategias para estimar dinámicamente el ancho de los *kernels* gaussianos utilizados en los modelados, permitiendo obtener resultados satisfactorios independientemente de las características de cada secuencia. En el caso del modelado del fondo se ha propuesto una eficiente estrategia basada en el análisis estadístico de las diferencias entre píxeles de imágenes consecutivas. En el caso del modelado del primer plano se ha diseñado una estrategia basada en *Mean-Shift* que, agrupando los píxeles de referencia en regiones homogéneas, permite estimar con gran calidad el ancho más adecuado para modelar cada región.
- **Actualización de las posiciones de los objetos previamente detectados:** Por otro lado, gracias a la aplicación de una innovadora estrategia de seguimiento, basada en un filtro de partículas específicamente diseñado para trabajar con un número variable de regiones móviles (descrito en el apéndice C), se consigue actualizar las posiciones espaciales de los píxeles de referencia pertenecientes a regiones móviles previamente detectadas. Así, se ha conseguido mejorar la calidad del modelado del primer plano y se ha reducido muy notablemente su coste computacional asociado. Además, las predicciones proporcionadas por dicho filtro se han empleado para obtener información a priori que se ha utilizado en un clasificador bayesiano que ha sido específicamente diseñado para poder hacer uso de cualquier tipo de probabilidad a priori de la que se disponga.

El resultado de la aplicación de todas estas mejoras es un sistema de detección que proporciona resultados de gran calidad en un amplio número de secuencias con contenido crítico para la detección de objetos móviles, mejorando la calidad de las detecciones obtenidas en estrategias paramétricas como la propuesta en el capítulo 4 y en otras estrategias basadas en el modelado no paramétrico del fondo y del primer plano.

- **Estrategias para reducir las falsas detecciones y para mejorar la eficiencia computacional de los métodos de detección basados en el modelado no paramétrico:** Aunque el sistema descrito en el capítulo 5 mejora la calidad de los resultados del resto de estrategias con las que se ha comparado, se ha observado que presenta algunas limitaciones: tiene asociado un elevado coste computacional y da lugar a un elevado número de falsas detecciones debidas a las sombras y a los reflejos provocados por los objetos móviles en sus desplazamientos. Es por eso que, en el capítulo 6, se han propuesto algunas estrategias que permiten reducir estas limitaciones:
  - **Aplicación de un innovador conjunto de características de apariencia:** Como alternativa a las comúnmente utilizadas componentes de color *RGB* se ha propuesto la utilización de las componentes de color normalizado y el módulo del gradiente de la saturación de los píxeles. De ese modo se ha conseguido reducir muy notablemente la influencia de las sombras y los reflejos en las detecciones, obteniéndose valores de *Precision* muy superiores a los obtenidos con otras estrategias que utilizan las componentes *RGB* de los píxeles.
  - **Información de posición únicamente en el modelado del primer plano:** Suprimiendo la información espacial del modelado del fondo se consiguen importantes reducciones del coste computacional asociado a dicho modelado y, conse-

cuentemente, se consigue mejorar la eficiencia computacional del sistema de detección.

- **Utilización de máscaras de Regiones de Interés:** Para conseguir un ahorro computacional aún mayor se han utilizado máscaras de Regiones de Interés que determinan qué píxeles deben ser analizados en cada instante. Dichas máscaras se obtienen como resultado de una ingeniosa combinación de: regiones resultantes del filtro de partículas utilizado para seguir a los objetos móviles previamente detectados; y regiones resultantes de un muestreo aleatorio que hemos denominado *Windowed Random Sampling* (*WRS*) y que proporciona pequeñas regiones aleatoriamente distribuidas en ventanas uniformemente repartidas en cada imagen. Con las primeras se consigue tener en cuenta los píxeles en los que es más probable la presencia de objetos móviles que ya han sido detectados anteriormente, mientras que con el *WRS* se cubre la posible aparición de nuevos objetos en la escena.

Como resultado de la aplicación de estas estrategias se ha mejorado tanto la calidad de los resultados (principalmente por la supresión de las falsas detecciones debidas a sombras y reflejos) como el coste computacional asociado al sistema (con porcentajes de mejora que van desde el 37 % hasta el 99 %, en función del contenido móvil presente en cada secuencia).

7.2. Trabajo futuro

En el campo relacionado con la segmentación temporal de secuencias se consideran distintas alternativas de trabajo futuro para que, partiendo del sistema descrito en el capítulo 3, sea posible obtener segmentaciones más precisas que ofrezcan mayor número de posibilidades en cuanto a la generación de resúmenes y a la extracción de información relevante de las secuencias. Mientras que el sistema propuesto se ha centrado en la segmentación temporal de secuencias a partir de la detección de cambios de toma, la división temporal de vídeos también puede llevarse a cabo a nivel de escenas o de planos, obteniéndose otros tipos de segmentación. Por lo tanto, una de las posibilidades de evolución del sistema propuesto podría ser la división de las secuencias considerando estos niveles de clasificación:

- **Segmentación en planos:** Por un lado se puede obtener una segmentación más precisa, ya que, como se ha descrito en la sección 2.1.5 del capítulo 2, es posible la división de una toma en planos. Así, una vez detectadas las tomas de las que se compone un vídeo, se puede analizar el movimiento global dentro de cada una y así identificar cuándo se produce alguna de las transiciones asociadas a los cambios de plano (*pan*, *tilt*, *zoom* o *traveling*), ofreciéndose la posibilidad de generar resúmenes más completos.
- **Segmentación en escenas:** Por otro lado, una escena está constituida por una o más tomas en las que todas las imágenes tienen relación con uno o varios objetos en un mismo entorno. Por lo tanto, si se consigue caracterizar de alguna forma el conjunto de objetos móviles que aparece a lo largo de las tomas detectadas, las tomas consecutivas que contengan información similar podrían unirse en escenas, las cuales permiten la generación de resúmenes atendiendo al contenido del vídeo y no únicamente a los cambios o variaciones que se produzcan en las cámaras.

Por otro lado, en el campo de la detección de objetos móviles se ha comprobado que las estrategias basadas en el modelado no paramétrico del fondo y el primer plano son capaces de proporcionar resultados mucho mejores que las basadas en mezclas de gaussianas. Sin embargo, a pesar de las grandes mejoras computacionales logradas con las estrategias propuestas en el capítulo 6, su coste computacional sigue siendo muy superior al de las estrategias paramétricas como la propuesta en el capítulo 4. Además, otra de las opciones que no se ha explotado y que tal vez podría dar lugar a mejores resultados es la de utilizar más características de apariencia de los píxeles como, por ejemplo, su profundidad en la escena. Teniendo en cuenta estas consideraciones se proponen las siguientes alternativas para continuar con el trabajo realizado en el campo de la detección de objetos:

- **Implementación en una *GPGPU*:** Para obtener una implementación eficiente del sistema de detección propuesto en el capítulo 5, haciendo uso de las estrategias de mejora descritas en el capitulo 6, se propone programarlo en una unidad de procesamiento gráfico de propósito general (*General-Purpose Graphical Processing Unit*, *GPGPU*) (Mohanty, 2009). Estos procesadores, gracias a su alto rendimiento y a su gran capacidad para paralelizar operaciones (realizan cientos, o incluso miles de operaciones en paralelo), cada vez están más presentes en dispositivos fijos y móviles (ordenadores, *smart-cameras*, etc.) (Tsai et al., 2011). Dado que en los esquemas de detección como los presentados en esta tesis se realizan las mismas operaciones sobre cada píxel de cada imagen, utilizando *GPGPUs* es posible realizar en paralelo las operaciones correspondientes a cada píxel. De este modo se pueden obtener importantes mejoras computacionales que permitan la utilización de este tipo de estrategias en aplicaciones en las que trabajar en tiempo real es un requisito imprescindible.
- **Incorporación de información de profundidad:** En las estrategias de detección descritas a lo largo de la tesis se ha utilizado tanto información de color de los píxeles (*RGB* en los capítulos 4 y 5 y color normalizado en el capítulo 6) como, en algunos casos, su posición dentro de la imagen (capítulos 5 y 6) y el gradiente de su saturación (capítulo 6). Sin embargo, en los últimos años han aparecido numerosos modelos de cámaras que, además de información de color, proporcionan información de profundidad de los píxeles (Foix et al., 2011). Utilizando esta información, junto con la propuesta en las estrategias descritas a lo largo de la tesis, es de esperar que sea posible reducir la cantidad de falsas detecciones debidas a las sombras y reflejos, así como las resultantes de cualquier cambio de iluminación. Sin embargo, también es posible que no pueda ser utilizada en cualquier tipo de escenario, ya que la precisión de la profundidad depende de la distancia de los objetos a la cámara, siendo peor a medida que esta distancia aumenta (Reynolds et al., 2011). Por lo tanto, puede ser interesante analizar qué mejoras se obtienen añadiendo esta información de profundidad a los esquemas de modelado propuestos y en qué escenarios y condiciones puede ser utilizada efectivamente.

Por último, se plantea la posibilidad de desarrollar un sistema de generación de resúmenes que utilice tanto los resultados de la estrategia de segmentación temporal descrita en el capítulo 3, como las detecciones resultantes de la aplicación de cualquiera de las estrategias propuestas en los capítulos 4, 5 y 6. Dicho sistema constaría de tres etapas. La primera sería la encargada de llevar a cabo la segmentación temporal en tomas de las secuencias.

La segunda se centraría en el análisis del movimiento dentro de cada una de las tomas previamente obtenidas. La tercera y última, partiendo de la información obtenida de la segmentación temporal (inicio y fin de las tomas, duración de las mismas, etc.) y de las características de los objetos móviles detectados en cada toma (número de objetos, su tamaño, su velocidad de desplazamiento, etc.), ofrecería la posibilidad de generar resúmenes atendiendo no sólo a criterios temporales, sino a las características de los objetos móviles que aparezcan en cada toma.

Apéndice A

Descripción de las bases de datos utilizadas

La vida es el arte de sacar conclusiones suficientes
a partir de datos insuficientes.
Samuel Butler (1835-1902),
novelista inglés.

Resumen: En este apéndice se describen las bases de datos utilizadas para llevar a cabo la evaluación de las estrategias de detección descritas a lo largo de esta tesis. En el caso de las estrategias para la detección de cambios de toma se ha hecho uso de una base de datos constituida por más de 30 secuencias, cuya duración supera las 3 horas y en las que es posible localizar más de 1000 transiciones entre tomas. Para evaluar las estrategias de detección de objetos móviles se ha utilizado una base de datos compuesta por 13 secuencias con contenido especialmente crítico para dichas estrategias como, por ejemplo: sombras y reflejos, cambios de iluminación, fondos dinámicos, vibraciones de la cámara y objetos móviles parecidos a regiones del fondo.

A.1. Base de datos para la segmentación temporal de secuencias

Para evaluar la calidad del sistema de segmentación temporal presentado en el capítulo 3 se ha utilizado una base de datos compuesta por más de 30 secuencias de vídeo, las cuales contienen más de 1000 tomas y cuya duración total aproximada es de 3 horas.

La mayor parte de estas secuencias son dibujos animados, musicales y reportajes, debido a la alta complejidad que supone la gran cantidad de movimiento en este tipo de vídeos, así como por el alto contenido de transiciones que poseen. En la tabla A.1 se muestra un resumen con las principales características de las secuencias utilizadas, agrupadas por categorías. Algunas imágenes representativas de algunas de estas secuencias se pueden ver en la figura A.1.

Figura A.1: Imágenes representativas de algunas de las secuencias de la base de datos utilizada para evaluar la estrategia de detección de cambios de toma.

Tipo de secuencia	Duración (seg.)	Nº de imágenes	Nº de trans. abruptas	Nº de trans. graduales	Nº total de transiciones
Dibujos	904	22640	279	7	286
Musicales	1244	33112	206	8	214
Reportajes	1589	40629	82	77	159
Otros	7012	171255	284	1	285
Total	10749	267636	851	93	944

Tabla A.1: Resumen de secuencias utilizadas, con sus duraciones y el número de transiciones abruptas y graduales que poseen.

Normalmente, las transiciones abruptas son mucho más frecuentes que las graduales (en la base de datos utilizada se pueden localizar aproximadamente 9 transiciones abruptas por cada transición gradual). Sin embargo, tal y como se puede ver en los datos mostrados en la tabla A.1, en las secuencias de tipo reportaje el número de transiciones graduales es muy superior (similar al de transiciones abruptas), motivo por el cual se han seleccionado estas secuencias para probar la calidad de la estrategia desarrollada para detectar este tipo de transiciones.

Por otro lado, además de por su gran cantidad de movimiento, la cual supone una gran dificultad a la hora de evitar la detección de falsos cambios de toma, los dibujos animados y los musicales se han seleccionado debido a su alta frecuencia de cambios de toma (en media, más de 13 transiciones por minuto), muy superior a la de otros tipos de secuencias.

Figura A.2: Imágenes representativas de las secuencias de la base de datos utilizada para evaluar las estrategias de detección de objetos móviles.

A.2. Base de datos para la detección de objetos móviles

Para evaluar la calidad de la estrategias propuestas en los capítulos 4, 5 y 6 se han utilizado 13 secuencias pertenecientes a distintas bases de datos. Las características más relevantes de estas secuencias, que hacen que sean adecuadas para la evaluación de las estrategias de detección de objetos móviles son:

- Presencia de sombras y reflejos provocados por los objetos móviles.
- Objetos móviles con regiones muy parecidas al fondo.
- Objetos móviles que se quedan parados distintos periodos de tiempo.
- Vibraciones de la cámara de grabación.
- Cambios de iluminación que afectan tanto a toda la imagen como únicamente a algunas regiones de la misma.
- Modificación prolongada de objetos estáticos que forman parte del fondo.
- Fondos con elementos no estáticos.

Nombre	Num. imágenes	Duración (seg.)	Dimensiones (ancho×alto)	Detección manual	Porcentaje movimiento	Num. objetos	Base de datos
Lab_001	325	13	288×352	Parcial	1,00	1	UPM-GTI
Lab_002	380	15	288×352	Parcial	2,47	2	UPM-GTI
Lab_003	550	22	198×272	Parcial	2,72	1	UPM-GTI
Lab_004	500	20	196×282	Parcial	2,79	1	UPM-GTI
Lab_005	250	10	288×352	Parcial	0,94	2	UPM-GTI
Lab_006	700	28	167×266	Parcial	5,25	1	UPM-GTI
Pets_001	1452	58	288×384	Parcial	1,12	8	PETS
Pets_002	500	20	288×384	Total	0	0	PETS
Pets_003	435	17	288×384	Parcial	0,37	4	PETS
Pets_004	795	31	288×360	Parcial	5,40	26	PETS
Pets_005	795	31	288×360	Parcial	6,27	27	PETS
Wall_001	293	11	120×160	Total	6,79	1	Wallflower
Wall_002	287	11	120×160	Total	1,52	1	Wallflower
Total	7262	287	-	-	2,82	-	-

Tabla A.2: Secuencias utilizadas para evaluar la calidad de las estrategias de detección de objetos móviles.

Para obtener datos cuantitativos de la calidad obtenida en las detecciones se ha llevado a cabo una detección manual sobre todas las secuencias, la cual se ha utilizado como detección de referencia para comparar con los resultados obtenidos con las distintas estrategias propuestas. Esta detección, dependiendo de las características de cada secuencia, se ha llevado a cabo en todas las imágenes o únicamente en algunas de ellas. En la sección A.2.1 se ofrecen más detalles relativos al modo de obtención de estas detecciones de referencia.

En la figura A.2 se muestra una imagen representativa de cada una de las secuencias utilizadas y en la tabla A.2 se muestra un resumen con la información más relevante de las mismas. Dicha información se compone de los siguientes campos:

- Nombre: Muestra el nombre de cada secuencia.
- Num. imágenes: Indica el número de imágenes de cada secuencia.
- Duración: Indica la duración, en segundos, de las secuencias.
- Detección manual: Indica el tipo de detección de referencia del que se dispone:
  - Total: Cuando la detección manual se ha realizado sobre todas las imágenes de la secuencia.
  - Parcial: Cuando sólo se ha llevado a cabo sobre algunas de las imágenes.
- Porcentaje movimiento: Porcentaje medio de píxeles móviles.
- Num. objetos: Número de objetos móviles que aparecen en la secuencia.
- Base de datos: Base de datos a la que pertenecen las secuencias, las cuales pueden ser:
  - UPM-GTI: Nuestra propia base de datos, compuesta por secuencias grabadas en un laboratorio.
  - PETS: Secuencias de la base de datos PETS (Computational Vision Group).

- Wallflower: Secuencias de la base de datos Wallflower (Toyama et al., 1999).

Como ya se ha dicho anteriormente, las 6 secuencias de la base de datos que hemos denominado *UPM-GTI* han sido grabadas en un laboratorio. Algunas de las características más relevantes de estas secuencias, que hacen que sean útiles para evaluar la calidad de las detecciones logradas, son las siguientes:

- Todas ellas contienen una gran cantidad de sombras y reflejos provocados por los objetos móviles y, además, los objetos móviles poseen regiones muy parecidas a algunas zonas del fondo.
- En algunos casos se pueden apreciar pequeñas vibraciones de la cámara con la que fueron grabadas.
- En *Lab*_003 y en *Lab*_006 aparecen objetos móviles que se quedan totalmente estáticos: durante un periodo breve de tiempo en el caso de *Lab*_006 y durante un periodo mucho mayor en el caso de *Lab*_003.
- En *Lab*_004 un objeto móvil modifica permanentemente parte del fondo del escenario.

Por otro lado, las secuencias de la base de datos *PETS* se corresponden con distintos escenarios al aire libre en los que las principales dificultades para la detección son:

- Todas ellas contienen numerosos elementos dinámicos en el fondo, principalmente vegetación agitada por el viento.
- Al igual que en el caso de las secuencias de la base de datos *UPM-GTI*, también contienen sombras y reflejos provocados por los objetos móviles en sus desplazamientos.
- *Pets*_001 contiene numerosos objetos móviles de muy distintos tamaños y, a lo largo de la misma, algunos de esos objetos pasan a formar parte del fondo.
- *Pets*_002 no contiene ningún objeto móvil, pero muestra numerosos cambios de iluminación que afectan con distinto grado de intensidad a distintas zonas del escenario.
- *Pets*_004 y *Pets*_005 muestran objetos móviles desde su comienzo hasta su final y, además, se caracterizan por un bajo contraste entre dichos objetos y los fondos de las secuencias.

Por último, las dos secuencias de la base de datos *Wallflower* se caracterizan por su fondo altamente dinámico y, en el caso de *Wall*_001 por las sombras y reflejos que provoca el objeto móvil que aparece en la secuencia.

A.2.1. Generación de las detecciones de referencia

Normalmente, las detecciones de referencia (*ground truth*) son esenciales para llevar a cabo un análisis cuantitativo de la calidad de los resultados proporcionados por las estrategias de detección de objetos móviles (Elhabian et al., 2008). Estas detecciones de referencia se pueden obtener de varias formas (Rosin, 1998): a partir de datos sintéticos, mediante anotaciones manuales, o a partir de las opiniones subjetivas de múltiples observadores humanos. Nosotros, hemos decidido hacer uso de detecciones de referencia basadas en anotaciones manuales.

Aunque, en principio, puede parecer que anotando manualmente los objetos móviles presentes en cada imagen es posible obtener una única detección de referencia, se debe

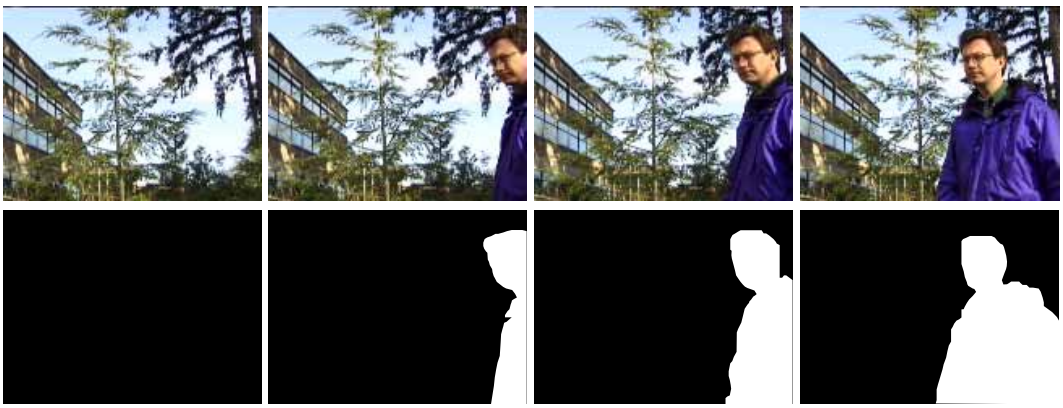


Figura A.3: Ejemplo de detecciones de referencia en el caso de una secuencia con objetos no estáticos en el fondo.

Figura A.4: Ejemplo de detecciones de referencia en el caso de una secuencia en la que un objeto del fondo pasa a ser móvil.

considerar que anotadores diferentes pueden interpretar el movimiento de las secuencias de forma distinta y, por lo tanto, pueden dar lugar a distintas detecciones de referencia (Tan et al., 2002). Es más, dos anotaciones de una misma secuencia, realizadas por un mismo anotador, pueden ser distintas, ya que dicho anotador puede realizar distintas interpretaciones del movimiento de los objetos. Nosotros, para tratar de reducir estas posibles diferencias hemos establecido las siguientes pautas de anotación:

- Cualquier objeto del fondo que se mueva por causas naturales (objetos movidos por el viento, olas del mar, lluvia, etc.) se debe etiquetar como parte del fondo. En la figura A.3 se pueden observar algunas de las anotaciones para el caso de un secuencia con movimiento de este tipo en el fondo (un árbol es agitado fuertemente por el viento).

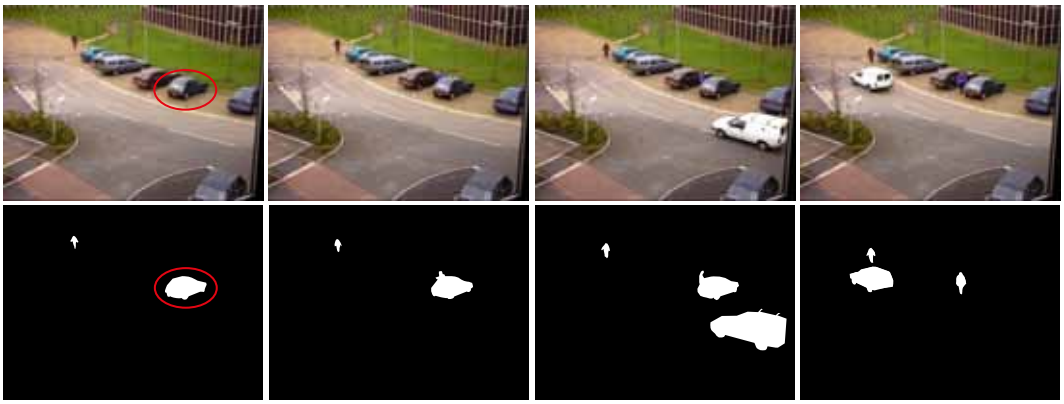


Figura A.5: Ejemplo de detecciones de referencia en el caso de una secuencia en la que un objeto móvil se queda parado.

- Cualquier objeto del fondo que comience a moverse por causas no naturales debe etiquetarse como parte del primer plano desde el momento en que comience su movimiento (no hay que aplicar ningún tiempo mínimo para garantizar que dicho objeto ha pasado a ser parte del contenido móvil de la secuencia). En la figura A.4 se muestra un ejemplo de esta situación en el caso de una secuencia en la que un objeto del fondo (un monitor) es sustraido por un individuo.
- Si un objeto móvil se queda parado seguirá considerándose móvil durante 150 imágenes, lo que equivale a unos 6 segundos, antes de ser considerado como parte del fondo. De este modo se garantiza que un objeto móvil realmente se ha quedado parado, evitando la pérdida de detecciones en los frecuentes casos en los que los objetos móviles interrumpen sus desplazamientos durante breves instantes de tiempo. En la figura A.5 se han representado las detecciones de referencia en el caso de una secuencia en la que un coche acaba de detenerse (rodeado por un círculo rojo en la primera columna de imágenes, en las que se muestra el instante aproximado de su detención). En la segunda y tercera columnas de imágenes, correspondientes a instantes temporales que no superan las 150 imágenes previamente mencionadas, a pesar de estar parado, el coche sigue siendo etiquetado como parte del primer plano. Sin embargo, en la cuarta columna de imágenes, en la que ya se han superado las 150 imágenes, se puede observar que ha sido etiquetado como parte del fondo.

Por último, hay que señalar que debido al excesivo trabajo que supone el etiquetado de más de 7000 imágenes, no se han etiquetado todas. Las secuencias en las que no hay contenido móvil ($Pets\_002$) y aquellas en las que el número de imágenes con contenido móvil es pequeño ($Wall\_001$ y $Wall\_002$) han sido etiquetadas en su totalidad. Sin embargo, en el resto de secuencias se ha etiquetado únicamente una imagen de cada 25, empezando por la primera imagen de cada secuencia. De este modo se ha conseguido que las detecciones de referencia sean representativas de cada una de las secuencias analizadas de principio a fin.

Apéndice B

Probabilidad acumulada de una gaussiana multidimensional en función de la distancia a su centro

Cuando no se puede lograr lo que se quiere,
mejor cambiar de actitud.

Publio Terencio Afer (195 AC-159 AC),
autor cómico latino.

B.1. Introducción

En ocasiones es interesante determinar cuál es la distancia mínima al centro de una distribución gaussiana multidimensional que acumula un porcentaje concreto de la probabilidad de dicha distribución. Este cálculo, llevado a cabo directamente sobre la distribución gaussiana original puede resultar excesivamente tedioso y, por ese motivo, es necesario transformar dicha distribución en otra más sencilla de analizar. En este apéndice se describe el modo en el que se puede llevar a cabo esta transformación. En primer lugar se analiza el caso genérico de una gaussiana multidimensional para, finalmente, concretar con el resultado correspondiente a una gaussiana bidimensional.

B.2. Análisis de una gaussiana multidimensional

Relacionar la probabilidad que acumula una distribución gaussiana multidimensional definida por D variables aleatorias correlacionadas, $\{X_1, X_2 \dots X_D\}$, en función de la distancia a su centro, es excesivamente complicado. Es por eso que resulta de gran utilidad aplicar las transformaciones que se describen a continuación para pasar de este tipo de gaussianas a otro tipo de funciones que faciliten este análisis. En primer lugar se muestra que mediante la aplicación de una transformación afín es posible pasar de una distribución gaussiana multidimensional genérica a otra gaussiana multidimensional estándar (orientada con los ejes de

coordenadas y con la misma desviación típica en cada una de sus componentes), centrada en el origen. Obtenida esta nueva gaussiana, se analiza su relación con una distribución unidimensional de tipo *chi-cuadrado*, en la que la relación entre probabilidad acumulada y distancia a su origen es mucho más sencilla.

Sea un conjunto de N muestras D-dimensionales, $\{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^D$, pertenecientes a una función densidad de probabilidad gaussiana que se define como:

$$f_X(\mathbf{x}; \mu_X, \Sigma_X) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_X|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_X)^T \Sigma_X^{-1} (\mathbf{x}-\mu_X)\right) \tag{B.1}$$

donde μ_X es el vector que determina el centro de la gaussiana y Σ_X es su matriz de escala. Aplicando sobre este conjunto de muestras una transformación afín definida como $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, donde $\mathbf{A}$ es una matriz y $\mathbf{b}$ es un vector, es posible obtener muestras de cualquier otra distribución gaussiana D-dimensional. Según el teorema fundamental de transformación de variables aleatorias (Papoulis y Pillai, 2002), la media, μ_Y, y matriz de escala, Σ_Y, de la distribución resultante son:

$$\begin{aligned} \mu_Y &= \mathbf{b} + \mathbf{A}\mu_X \\ \Sigma_Y &= \mathbf{A}\Sigma_X\mathbf{A}^T \end{aligned} \tag{B.2}$$

Utilizando este resultado podemos diseñar una transformación afín que genere una gaussiana D-dimensional con media $\mu_Y = 0$ y matriz de escala $\Sigma_Y = \sigma^2\mathbf{I}$ (con varianza σ^2 en cada componente). Dado que Σ_X es una matriz real, simétrica y definida positiva, es posible descomponerla en sus autovectores, $\mathbf{V}$, y autovalores, Λ: $\Sigma_X = \mathbf{V}\Lambda\mathbf{V}^{-1}$, donde $\mathbf{V}$ es una matriz ortogonal ($\mathbf{V}^{-1} = \mathbf{V}^T$) y Λ es una matriz diagonal con valores estrictamente positivos. Por lo tanto, los parámetros de la anterior transformación afín son:

$$\begin{aligned} \mathbf{A} &= \sigma(\Lambda^{\frac{1}{2}})^{-1}\mathbf{V}^{-1} \\ \mathbf{b} &= -\sigma(\Lambda^{\frac{1}{2}})^{-1}\mathbf{V}^{-1}\mu_X \end{aligned} \tag{B.3}$$

Quedando dicha transformación definida como:

$$\mathbf{y} = \sigma(\Lambda^{\frac{1}{2}})^{-1}\mathbf{V}^{-1}(\mathbf{x}-\mu_X) \tag{B.4}$$

Geométricamente esto se puede interpretar como tres operaciones concatenadas: una traslación al origen (μ_X), una rotación que orienta la gaussiana con los ejes coordenados ($\mathbf{V}^{-1}$) y un escalado en cada eje ($\sigma(\Lambda^{\frac{1}{2}})^{-1}$).

Descrito el modo en el que es posible convertir una gaussiana multidimensional cualquiera a una gaussiana multidimensional estándar, a continuación se analiza el modo en el que dicha gaussiana estándar puede relacionarse con una función de tipo *chi-cuadrado*.

Sea un conjunto de variables aleatorias independientes, $\{Y_1, Y_2 \dots Y_D\}$, cuya función densidad de probabilidad es gaussiana y que, además, tienen un valor medio igual a 0 y

una varianza idéntica e igual a σ^2. Una variable aleatoria definida como la suma de los cuadrados de este conjunto de variables gaussianas,

$$Z = \sum_{i=1}^{D} Y_i^2 \tag{B.5}$$

posee una función densidad de probabilidad de tipo *chi-cuadrado* (Proakis, 2001):

$$f_Z(\mathbf{z}) = \frac{1}{\sigma^D 2^{\frac{D}{2}} \Gamma(\frac{1}{2}D)} \mathbf{z}^{\frac{D}{2}-1} \exp\left(-\frac{\mathbf{z}}{2\sigma^2}\right), \quad \mathbf{z} \geq 0 \tag{B.6}$$

donde $\Gamma(\cdot)$ es la función gamma y cuya función de distribución para valores pares de D es:

$$F_Z(\mathbf{z}) = P(Z \leq \mathbf{z}) = 1 - \exp\left(-\frac{\mathbf{z}}{2\sigma^2}\right) \sum_{k=0}^{\frac{D}{2}-1} \frac{1}{k!} \left(\frac{\mathbf{z}}{2\sigma^2}\right)^k, \quad \mathbf{z} \geq 0 \tag{B.7}$$

Por lo tanto, es posible relacionar la función de distribución de una gaussiana multidimensional genérica con la de una gaussiana estándar y esta, a su vez, con la de una *chi-cuadrado*:

$$P\left(\sigma^2(\mathbf{x}-\mu_X)^T\Sigma_X^{-1}(\mathbf{x}-\mu_X) \leq M\right) = P(\|Y\|_2^2 \leq M) = P(Z \leq M) \tag{B.8}$$

donde $((\mathbf{x}-\mu_X)^T\Sigma_X^{-1}(\mathbf{x}-\mu_X))$ es el cuadrado de la distancia de *Mahalanobis*, D_{Mah}, de la distribución inicial, $\|Y\|_2$ es la distancia euclídea de la distribución gaussiana estándar y M es un valor de probabilidad cualquiera. De este modo, obteniendo la probabilidad de la función *chi-cuadrado*, acumulada por debajo de M, se podrá calcular la distancia de *Mahalanobis* de la distribución inicial por debajo de la cual se encuentra acumulado ese mismo valor de probabilidad.

B.3. Análisis de una gaussiana bidimensional

Es esta sección, debido a que ha sido utilizado a lo largo de esta tesis, se analiza el resultado obtenido en la sección anterior, particularizado para una distribución gaussiana bidimensional. En este caso, la función densidad de probabilidad de una *chi-cuadrado* es de la forma:

$$f_Z(\mathbf{z}) = \frac{1}{\sigma^2} \exp\left(-\frac{\mathbf{z}}{2\sigma^2}\right), \quad \mathbf{z} \geq 0 \tag{B.9}$$

Si se define una nueva variable:

$$R = \sqrt{Z} \tag{B.10}$$

haciendo un cambio de variable sencillo sobre esta densidad de probabilidad se obtiene la función densidad de probabilidad de una *Rayleigh* (Taub y Schilling, 1986):

$$f_R(\mathbf{r}) = \frac{\mathbf{r}}{\sigma^2} \exp\left(-\frac{\mathbf{r}^2}{2\sigma^2}\right), \quad \mathbf{r} \geq 0 \tag{B.11}$$

cuya función de distribución es:

$$F_R(\mathbf{r}) = P(R \leq \mathbf{r}) = 1 - \exp\left(-\frac{\mathbf{r}^2}{2\sigma^2}\right), \quad \mathbf{r} \geq 0 \tag{B.12}$$

En el caso en el que $\sigma = 1$, esta función de distribución se relaciona con la de la gaussiana multidimensional inicial mediante la expresión:

$$P\left(\sqrt{(\mathbf{x} - \mu_X)^T \Sigma_X^{-1} (\mathbf{x} - \mu_X)} \leq M\right) = P(R \leq M) = 1 - \exp\left(-\frac{\mathbf{r}^2}{2}\right) \tag{B.13}$$

Por lo tanto, la distancia de *Mahalanobis*, D_{Mah}, que acumula un porcentaje P_0 de probabilidad de una gaussiana bidimensional se obtiene como:

$$D_{Mah} = \sqrt{-2\ln(1 - P_0)} \tag{B.14}$$

por lo que, si lo que se desea es acumular aproximadamente el 99 % de la probabilidad de una distribución gaussiana bidimensional (requisito que se ha impuesto en algunos de los algoritmos expuestos en distintas secciones de esta tesis), se deberán considerar las muestras cuya distancia de *Mahalanobis* al centro de dicha distribución sea menor que 3.

Apéndice C

Filtro de partículas para el seguimiento de múltiples regiones móviles

La estadística es una ciencia que demuestra
que si mi vecino tiene dos coches y yo ninguno,
los dos tenemos uno.

George Bernard Shaw (1856-1950),
escritor irlandés.

Resumen: En este apéndice se describe el filtro de partículas desarrollado para llevar a cabo el seguimiento de las múltiples regiones móviles, resultantes de los procesos de detección descritos en los capítulos 5 y 6, que permite obtener los desplazamientos aplicados en el modelado del primer plano descrito en la sección 5.5.1 y las probabilidades a priori descritas en la sección 5.5.2. Al contrario que los algoritmos de seguimiento tradicionales basados en filtros de partículas, el filtro desarrollado es capaz de trabajar de forma muy eficiente con un número variable de regiones móviles. En cada instante temporal las partículas se distribuyen entre todas las regiones móviles existentes y se realizan estimaciones independientes para cada región. Además, para detectar la presencia de nuevas regiones y asignarles partículas que permitan su seguimiento se ha elaborado una estrategia de análisis de regiones que permite identificar cuándo una región no está siendo tenida en cuenta por el filtro.

C.1. Introducción

El seguimiento de objetos móviles es una de las áreas más estudiadas dentro del campo del procesamiento de imágenes y es por eso que ha dado lugar a un extenso número de publicaciones en las que se proponen diversas estrategias para llevarlo a cabo (del Blanco et al., 2010).

Entre las estrategias de seguimiento más populares se encuentran aquellas basadas en

los métodos de Monte Carlo y su versión secuencial, comúnmente conocida como filtros de partículas (Isard y Blake, 1998) (Doucet et al., 2000) (Arulampalam et al., 2002).

La mayor parte de las propuestas que utilizan filtros de partículas para seguir múltiples regiones móviles (Meier y Ade, 1999) (Khan et al., 2004) (Serby et al., 2004) (Dore et al., 2007) (Smal et al., 2007) consideran un número fijo de regiones móviles para simplificar las complicadas tareas de inicialización, actualización y combinación de partículas. Algunas de estas propuestas tratan estas situaciones considerado que existe una única región a seguir, compuesta por múltiples sub-regiones (Serby et al., 2004) (Dore et al., 2007). Otras, como las propuestas en (Khan et al., 2004) y (Smal et al., 2007), proponen soluciones menos restrictivas: la primera utiliza un modelo dinámico que le permite trabajar con múltiples regiones móviles que interacionan entre sí, mientras que la segunda resuelve algunas situaciones de oclusión entre las regiones móviles que trata de seguir. Sin embargo, en el análisis de secuencias como las descritas a lo lago de esta tesis, en las que el número de objetos móviles es variable, los resultados proporcionados por estas estrategias no son adecuados.

Para obtener resultados satisfactorios en estas situaciones se ha desarrollado una novedosa estrategia de seguimiento, basada en un filtro de partículas específicamente diseñado para de seguir eficientemente un número variable y desconocido de regiones móviles. Para tal propósito el filtro propuesto asocia grupos de partículas a cada región a seguir y realiza estimaciones independientes para cada una de ellas. Además, mediante el análisis de la relación entre las partículas y las regiones existentes, identifica cuándo alguna de las regiones existentes no está siendo adecuadamente representada por las partículas y, de ese modo, permite que el filtro se adapte para tenerla en cuenta.

C.2. Descripción de la estrategia

El algoritmo propuesto está constituido por las etapas típicas de un filtro de partículas de tipo SIR (*Sequential Importance Resampling*) (Arulampalam et al., 2002): predicción, actualización, normalización y remuestreo. El objetivo de dicho algoritmo es estimar, de forma iterativa, la función densidad de probabilidad de un vector de estado, $\mathbf{u}^n$, a partir de un conjunto de medidas, $\mathbf{v}^n$, obtenidas de la imagen I^n en el instante temporal n. Dicha función de densidad se define como:

$$p(\mathbf{u}^n|\mathbf{v}^n, \mathbf{v}^{n-1}, \ldots, \mathbf{v}^1) \propto p(\mathbf{v}^n|\mathbf{u}^n)p(\mathbf{u}^n|\mathbf{v}^{n-1}, \ldots, \mathbf{v}^1) \tag{C.1}$$

donde el factor $p(\mathbf{v}^n|\mathbf{u}^n)$ es la verosimilitud del vector $\mathbf{u}^n$ dado el conjunto de medidas $\mathbf{v}^n$ y el factor $p(\mathbf{u}^n|\mathbf{v}^{n-1}, \ldots, \mathbf{v}^1)$ es la distribución predicha a partir de las observaciones correspondientes a los instantes anteriores (Isard y Blake, 1998). Para obtener esta estimación se lleva a cabo la evaluación de un conjunto de N_s partículas definidas como $\{\mathbf{u}_i^n, \varpi_i^n\}_{i=1}^{N_s}$, en las que ϖ_k^i es el peso de la partícula i-ésima.

Para tal propósito se ha decidido representar las regiones móviles detectadas mediante elipses orientadas con los ejes de las imágenes, cuya velocidad se ha asumido además

constante. Por ello, cada elipse se representará mediante un vector de estado definido como:

$$\mathbf{u}_i^n = \left(h_i^n, w_i^n, \dot{h}_i^n, \dot{w}_i^n, a_i^n, b_i^n\right)^T \tag{C.2}$$

donde el par (h_i^n, w_i^n) representa la posición del centro de la elipse i-ésima, el par $(\dot{h}_i^n, \dot{w}_i^n)$ hace referencia a la velocidad de dicha elipse y el par (a_i^n, b_i^n) representa las dimensiones de sus ejes.

C.2.1. Gestión del número variable de objetos

Una de las contribuciones más significativas del filtro de partículas propuesto es su capacidad para adaptarse a un número variable de regiones móviles a lo largo de una secuencia de imágenes. Para tal propósito, en cada imagen, I^n, las N_s partículas utilizadas se reparten entre las M_s regiones móviles detectadas en dicho instante. Esta distribución se lleva a cabo en función de la importancia que se le asigna a cada región móvil, la cual se ha decidido que sea proporcional a la cantidad de medidas que las conforman. De ese modo la estimación será más precisa para los objetos que ocupen mayor espacio en la imagen.

Una vez efectuada esta distribución, cada grupo de partículas será utilizado, independientemente del resto de grupos, para estimar el desplazamiento de su región asociada. Por lo tanto, en cada instante temporal se estará haciendo uso de un filtro de partículas independiente para cada región, con un número de partículas asociadas que dependerá del tamaño de las regiones en dicho instante. Sin embargo, ya que el número total de partículas es el mismo para todas las imágenes, el coste computacional asociado al filtrado será aproximadamente constante.

En primer lugar, antes de proceder a distribuir las partículas del modo que se acaba de describir, es necesario identificar si existe alguna región nueva que no esté siendo evaluada por ninguno de los grupos de partículas y, en el caso de que exista, habrá que asignarle un grupo de partículas que la represente. Para llevar a cabo esta asignación, la estrategia propuesta sustituye una de las partículas existentes (la de menor peso asociado, dentro del grupo más numeroso de partículas) por una que represente a la nueva región detectada. A continuación se describe el modo en el que estas nuevas regiones son detectadas.

C.2.1.1. Detección de regiones nuevas

En cada nueva imagen, la presencia de regiones móviles nuevas se comprueba mediante la evaluación de gaussianas bidimensionales que permiten determinar si las medidas actuales están o no están siendo cubiertas por las partículas existentes. Estas gaussianas se definen como:

$$g(h, w|\mathbf{u}_i^n) = exp\left(-\frac{1}{2}\left(\frac{(h-h_i^n)^2}{(\sigma_i^n(1))^2} + \frac{(w-w_i^n)^2}{(\sigma_i^n(2))^2}\right)\right) \tag{C.3}$$

donde (h, w) son las coordenadas de los píxeles dentro de la imagen y σ_i^n es la desviación típica que determina el ancho de las gaussianas en cada dimensión.

En cada instante, el cálculo de estas gaussianas se lleva a cabo únicamente sobre los J_s píxeles de los que se compone el vector de medidas de la imagen actual, los cuales se corresponden con el conjunto de píxeles clasificados como móviles en dicha imagen:

$$\mathbf{v}^n = \{\mathbf{v}_m^n\}_{m=1}^{J_s} = \{(h, w) \ / \ Pr(\phi|\mathbf{x}^n) \geq T_\phi\} \tag{C.4}$$

siendo T_ϕ el umbral que determina la probabilidad mínima para clasificar un píxel como parte del primer plano de la secuencia.

Se considera que una medida está cubierta por una partícula si dicha medida está contenida en la región que acumula el 99 % de la probabilidad de alguna de las gaussianas. De ese modo, la desviación típica de estas gaussianas deberá establecerse como (apéndice B):

$$\sigma_i^n = \left(\frac{a_i^n}{3}, \frac{b_i^n}{3}\right) \tag{C.5}$$

Por lo tanto, para considerarse cubierta por una partícula, las medidas deben verificar que:

$$\sum_{j=1}^{2} \frac{(\mathbf{u}_i^n(j) - \mathbf{v}_m^n(j))^2}{(\sigma_i^n(j))^2} \leq 1 \tag{C.6}$$

En los casos en los que ninguna medida de las asociadas a una región móvil esté cubierta por las partículas existentes se determina que dicha región móvil es nueva.

Las partículas creadas para representar a las nuevas regiones se inicializan de la siguiente forma:

- Su posición inicial, (h, w), será la dada por la mediana de las posiciones, en cada componente espacial, de las medidas asociadas a la nueva región móvil.
- Su velocidad inicial, $(\dot{h}, \dot{w})$, se establecerá aleatoriamente en torno a cero.
- Sus dimensiones, (a, b), se inicializan como $a = 3\sigma_h$ y $b = 3\sigma_w$, donde σ_h y σ_w son las desviaciones típicas de las coordenadas de los píxeles de la nueva región móvil en cada dimensión, filas y columnas. Consecuentemente, la nueva elipse cubrirá la mayor parte de las medidas de la nueva región (apéndice B).

C.2.1.2. Desaparición y unión de regiones

Por otro lado, además de la aparición de nuevas regiones, también se han de considerar otras situaciones frecuentes como la desaparición o la unión de regiones.

En el primer caso, si una región deja de estar presente, todas sus partículas asociadas se repartirán entre el resto de regiones existentes (con valores iniciales aleatorios entorno a la estimación del vector de estado asociado a cada región).

En el segundo caso, si dos o más regiones se unen, el resultado será considerado como una nueva región y, por lo tanto, será necesaria la asignación de un nuevo grupo de partículas que se adapte a sus características (asignación realizada del modo expuesto al final de la sección C.2.1.1). Se debe tener en cuenta que estas situaciones no se producen únicamente

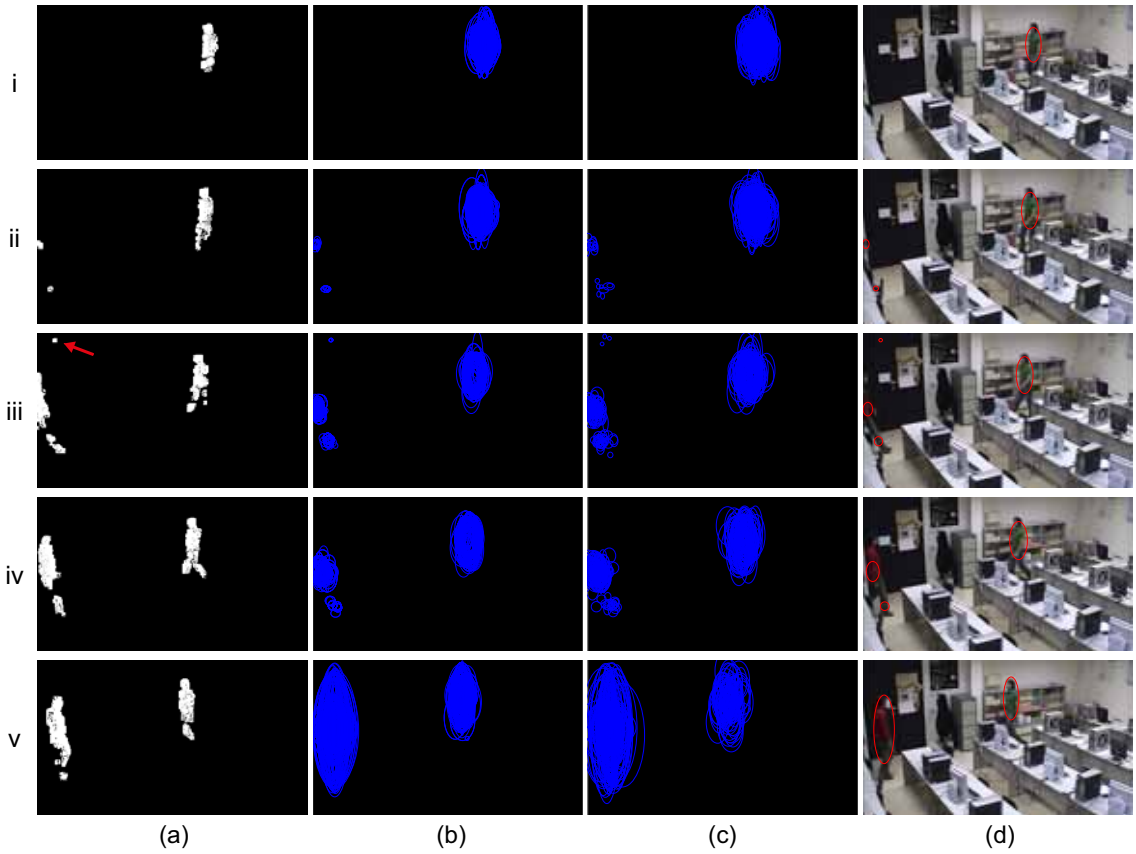


Figura C.1: Resultados obtenidos con el filtro de partículas propuesto. (a) Objetos móviles detectados (medidas). (b) Partículas asignadas a las regiones móviles. (c) Partículas predichas. (d) Estimación de los vectores de estado sobre las imágenes originales.

en las interacciones entre objetos móviles, sino que también tienen lugar cuando múltiples regiones de un mismo objeto móvil acaban uniéndose.

Para ilustrar todas estas situaciones, en la figura C.1 se han representado algunos resultados obtenidos con el filtro de partículas propuesto, aplicado a lo largo de una secuencia en la que inicialmente existe un único objeto móvil y, posteriormente, aparece un segundo objeto móvil. La primera columna de imágenes de esta figura contiene las detecciones obtenidas en distintos instantes de tiempo. La segunda columna muestra el conjunto de partículas asignadas a cada región móvil tras haberse aplicado la etapa de remuestreo. La tercera columna presenta las partículas resultantes de la etapa de predicción. Por último, la cuarta columna muestra las estimaciones de los vectores de estado, superpuestos sobre las imágenes originales. Se puede apreciar que en el momento en el que aparece el segundo objeto móvil (figura C.1.ii) se identifican dos nuevas regiones móviles y, consecuentemente aparecen dos nuevos grupos de partículas que permiten llevar a cabo el seguimiento de estas regiones. A lo largo de las imágenes siguientes se observa que, a medida que cada una de estas dos nuevas regiones crece, sus partículas asociadas se adaptan a su tamaño. Pasa-

do un tiempo (figura C.1.v), las dos regiones se unen y pasan a formar una única región representada mediante un único grupo de partículas adaptadas a sus características. En estos ejemplos también se puede apreciar que, debido a la presencia de una falsa detección (señalada con una flecha roja en la figura C.1.iii.a), ha aparecido un grupo de partículas que representa a una región que no debería estar siendo considerada. Sin embargo, en el siguiente instante representado en la figura (figura C.1.iv) dicha región ha dejado de ser detectada y, consecuentemente, sus partículas asociadas se han distribuido entre el resto de regiones móviles.

C.2.2. Evaluación de las partículas

Una vez repartidas las partículas entre todas las regiones móviles existentes, el siguiente paso consiste en asignar un valor de peso a cada partícula. Dicho valor deberá ser mayor para las partículas que mejor se adapten a la región móvil que tengan asociada, siendo menor a medida que peor representen a dicha región.

Sea una región móvil, $\{V_\tau^n\}_{\tau=1}^J$, constituida por J medidas. El peso de las partículas asociadas a esta región se calcula como:

$$\varpi_i^n = L(V_\tau^n|\mathbf{u}_i^n) \cdot \varpi_i^{n-1} \tag{C.7}$$

donde $L(V_\tau^n|u_i^n)$ es una función de verosimilitud que relaciona el conjunto de medidas de la región móvil con cada una de las partículas asociadas a esa región. Esta verosimilitud la hemos definido como:

$$L(V_\tau^n|\mathbf{u}_i^n) \propto \left(\beta\frac{J_{in}}{\pi a_i^n b_i^n} + (1-\beta)\frac{J_{in}}{J}\right)\prod_{\tau=1}^{J}(\rho + g(V_\tau^n|\mathbf{u}_i^n)Pr(\phi|V_\tau^n)) \tag{C.8}$$

donde J_{in} es el número de medidas cubiertas por la partícula u_i^n (según los criterios explicados en la sección C.2.1).

El primer factor de esta expresión, $F_1 = J_{in}/\pi ab$, representa la relación entre el número de medidas de la región móvil que están cubiertas por la partícula y el área de la partícula. Este factor dará lugar a mayores valores de verosimilitud cuanto más rellena de medidas esté la partícula.

El segundo factor, $F_2 = J_{in}/J$, representa la relación entre el número de medidas de la región móvil que cubre la partícula y el número total de medidas de la región. Por lo tanto, proporcionará mayores valores de verosimilitud cuantas menos medidas se deje fuera la partícula. La variable $\beta \in [0,1]$ se utiliza para determinar a cuál de estos dos factores se le da más importancia.

La figura C.2 muestra un ejemplo en el que aparece una región móvil y dos de sus partículas asociadas. En este ejemplo, la partícula de la imagen de la izquierda es más grande que la región que representa, por lo que el valor de F_1 al que da lugar es menor que el obtenido con la partícula de la imagen de la derecha, la cual está completamente llena de medidas. Sin embargo, si se evalúa el segundo factor, F_2, la partícula de la imagen de la izquierda contiene todas las medidas de la región, mientras que la de la imagen de la derecha

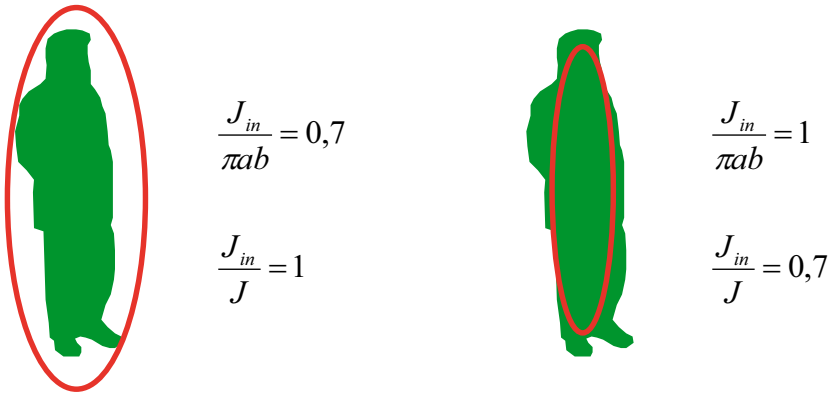


Figura C.2: Análisis de la calidad de las partículas en función de cómo se ajustan a la región de móvil que representan.

se deja fuera algunas de estas medidas. Consecuentemente, el valor de F_2 es mayor en el caso de la primera partícula. Si ambos factores tuvieran la misma influencia en la evaluación de las partículas, las dos partículas de este ejemplo supondrían el mismo aporte a la función de verosimilitud previamente definida. Es por eso que, si se desea favorecer uno u otro criterio (ajustar el tamaño de las partículas por exceso o por defecto), es necesario utilizar el factor β. En el caso de la estrategia de detección descrita en el capítulo 5 se ha comprobado que la calidad de los resultados obtenidos no sufre variaciones relevantes, indistintamente del valor asignado a β. No obstante, en el caso de la estrategia descrita en el capítulo 6, en la que el tamaño de las partículas influye en el área analizada en cada instante, se ha observado que con valores pequeños de β, al favorecerse la propagación de las partículas de mayor tamaño, se reduce ligeramente el número de píxeles móviles no detectados.

El tercero de los factores utilizados es $F_3 = \prod_{\tau=1}^{J} (\rho + g(\cdot)Pr(\cdot))$. Este factor se obtiene combinando la contribución de cada una de las medidas de la región, $g(\cdot)$, ponderada por la probabilidad de estas medidas de formar parte del primer plano, $Pr(\cdot)$ (obtenida como resultado de cualquiera de las estrategias de detección expuestas en los capítulos 5 y 6). En esta expresión, ρ es un parámetro de regulación que se utiliza para evitar la influencia de medidas con valores de probabilidad nulos. Cuanto mayor sea el valor de probabilidad de las medidas cubiertas por una partícula, mayor será el valor de F_3. De este modo se favorecerá a las partículas que se encuentren situadas en las zonas de las regiones móviles que más claramente han sido etiquetadas como parte del primer plano de la secuencia.

En último lugar, una vez obtenidos los pesos de las partículas, se procede a remuestrear las partículas en función de esos pesos (más réplicas para las partículas con mayor peso asociado) y, aplicando un modelo dinámico de velocidad constante, se predice su estado para la siguiente imagen.

Bibliografía

Y así, del mucho leer y del poco dormir, se le secó el celebro de manera que vino a perder el juicio.
Miguel de Cervantes Saavedra

Ahmed, M., Karmouch, A. y Abu-Hakima, S. Key frame extraction and indexing for multimedia databases. En *Vision Interface Conference, Trois-Rivières, Canada*, páginas 506–11. 1999.

Aires, K., Santana, A. y Medeiros, A. Optical flow using color information: preliminary results. En *Proceedings of the 2008 ACM symposium on Applied computing*, páginas 1607–1611. ACM, 2008.

Amato, A., Mozerov, M., Bagdanov, A. y Gonzalez, J. Accurate moving cast shadow suppression based on local color constancy detection. *Image Processing, IEEE Transactions on*, (99), páginas 1–1, 2011.

Arulampalam, M., Maskell, S., Gordon, N. y Clapp, T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Transactions on*, vol. 50(2), páginas 174–188, 2002.

Atev, S., Masoud, O. y Papanikolopoulos, N. Practical mixtures of Gaussians with brightness monitoring. En *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*, páginas 423–428. IEEE, 2005.

Ballard, D. Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, vol. 13(2), páginas 111–122, 1981.

Bayes, M. y Price, M. An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions*, vol. 53, página 370, 1763.

Benezeth, Y., Jodoin, P., Emile, B., Laurent, H. y Rosenberger, C. Review and evaluation of commonly-implemented background subtraction algorithms. En *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, páginas 1–4. IEEE, 2009.

BESCÓS, J. *Segmentación temporal de secuencias de vídeo*. Tesis Doctoral, Universidad Politécnica de Madrid, 2001.

BICEGO, M., CRISTANI, M. y MURINO, V. Unsupervised scene analysis: A hidden Markov model approach. *Computer vision and image understanding*, vol. 102(1), páginas 22–41, 2006.

DEL BLANCO, C. R., JAUREGUIZAR, F. y GARCÍA, N. Robust Tracking in Aerial Imagery Based on an Ego-Motion Bayesian Model. *EURASIP Journal on Advances in Signal Processing*, vol. 2010(30), páginas 1–18, 2010.

BOUGUET, J. Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. *Intel Corporation, Microprocessor Research Labs, OpenCV Documents*, 1999.

BOUTHEMY, P., GELGON, M. y GANANSIA, F. A unified approach to shot change detection and camera motion characterization. *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9(7), páginas 1030–1044, 1999.

BOUTTEFROY, P., BOUZERDOUM, A., PHUNG, S. y BEGHDADI, A. On the analysis of background subtraction techniques using Gaussian Mixture Models. En *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, páginas 4042–4045. IEEE, 2010.

BOUWMANS, T., EL BAF, F. y VACHON, B. Background modeling using mixture of gaussians for foreground detection-a survey. *Recent Patents on Computer Science*, vol. 1(3), páginas 219–237, 2008.

BREZEALE, D. y COOK, D. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems Man and Cybernetics-Part C-Applications Reviews*, vol. 38(3), páginas 416–430, 2008.

CAMARA-CHAVEZ, G., CORD, M., PHILIPP-FOLIGUET, S., PRECIOSO, F. y DE ALBUQUERQUE ARAUJO, A. Robust scene cut detection by supervised learning. En *Proceedings of EUSIPCO*. 2006.

CAO, J. y CAI, A. A robust shot transition detection method based on support vector machine in compressed domain. *Pattern Recognition Letters*, vol. 28(12), páginas 1534–1540, 2007.

CHAISORN, L., MANDERS, C. y RAHARDJA, S. Video retrieval - evolution of video segmentation, indexing and search. En *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, páginas 16 –20. 2009.

CHASANIS, V., LIKAS, A. y GALATSANOS, N. Simultaneous detection of abrupt cuts and dissolves in videos using support vector machines. *Pattern Recognition Letters*, vol. 30(1), páginas 55–65, 2009.

CHAUDHURI, D. y AGRAWAL, A. Split-and-merge Procedure for Image Segmentation using Bimodality Detection Approach. *Defence Science Journal*, vol. 60(3), página 290, 2010.

CHENG, Y. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17(8), páginas 790–799, 1995.

CIOCCA, G. A robust multi-feature cut detection algorithm for video segmentation. *ELCVIA: Electronic Letters on Computer Vision and Image Analysis*, vol. 9(1), páginas 32–46, 2010.

COMANICIU, D. y MEER, P. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24(5), páginas 603–619, 2002.

COMANICIU, D., RAMESH, V. y MEER, P. The variable bandwidth mean shift and data-driven scale selection. En *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, páginas 438–445. IEEE, 2001.

COMPUTATIONAL VISION GROUP, . Pets: Performance evaluation of tracking and surveillance. Disponible en `http://www.cvg.rdg.ac.uk/`. Computational Vision Group, University of Reading.

COTO, E. Métodos de segmentación de Imágenes Médicas. *Universidad central de Venezuela, Facultad de Ciencias*, 2005.

CRISTANI, M., FARENZENA, M., BLOISI, D. y MURINO, V. Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in Signal Processing*, vol. 2010, página 43, 2010.

CUCCHIARA, R., GRANA, C., PICCARDI, M. y PRATI, A. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, páginas 1337–1342, 2003.

CUCCHIARA, R., PRATI, A. y VEZZANI, R. Real-time motion segmentation from moving cameras. *Real-Time Imaging*, vol. 10(3), páginas 127–143, 2004.

CUEVAS, C., DEL BLANCO, C., GARCIA, N. y JAUREGUIZAR, F. Segmentation-tracking feedback approach for high-performance video surveillance applications. En *Image Analysis & Interpretation (SSIAI), 2010 IEEE Southwest Symposium on*, páginas 41–44. IEEE, 2010a.

CUEVAS, C. y GARCÍA, N. Moving object detection for real-time high-quality lightweight applications on smart cameras. En *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, páginas 479–480. IEEE, 2010a.

CUEVAS, C. y GARCÍA, N. Real-time shot detection based on motion analysis and multiple low-level techniques. En *Proceedings of SPIE*, vol. 7701, páginas 1–10. 2010b.

CUEVAS, C. y GARCÍA, N. Tracking-based non-parametric background-foreground classification in a chromaticity-gradient space. En *Image Processing (ICIP), 2010 17th IEEE International Conference on*, páginas 845–848. IEEE, 2010c.

CUEVAS, C. y GARCÍA, N. Automatic bandwidth estimation strategy for high-quality non-parametric modeling based moving object detection. En *Image Processing (ICIP), 2011 18th IEEE International Conference on*, páginas 1797–1800. IEEE, 2011.

CUEVAS, C., GARCÍA, N. y SALGADO, L. A new strategy based on adaptive mixture of gaussians for real-time moving objects segmentation. En *Proceedings of SPIE*, vol. 6811, páginas 1–12. 2008.

CUEVAS, C., MOHEDANO, R., JAUREGUIZAR, F. y GARCÍA, N. High-quality real-time moving object detection by non-parametric segmentation. *Electronics Letters*, vol. 46(13), páginas 910–911, 2010b.

CUFI, X., MUNOZ, X., FREIXENET, J. y MARTI, J. A review of image segmentation techniques integrating region and boundary information. *Advances in Imaging and Electron Physics*, vol. 120, páginas 1–39, 2003.

DAMGHANIAN, B., HASHEMI, M. y AKBARI, M. A novel fade detection algorithm on H. 264/AVC compressed domain. *Advances in Image and Video Technology*, páginas 1159–1167, 2006.

DE BRUYNE, S., DE NEVE, W., DE WOLF, K., DE SCHRIJVER, D., VERHOEVE, P. y VAN DE WALLE, R. Temporal video segmentation on H. 264/AVC compressed bitstreams. *Advances in Multimedia Modeling*, páginas 1–12, 2006.

DING, J., LI, M., HUANG, K. y TAN, T. Modeling complex scenes for accurate moving objects segmentation. *Computer Vision–ACCV 2010*, páginas 82–94, 2011.

DORE, A., REGAZZONI, C. y MUSSO, M. Map particle selection in shape-based object tracking. En *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 5, páginas V–341. IEEE, 2007.

DOSHI, A. y BORS, A. Smoothing of optical flow using robustified diffusion kernels. *Image and Vision Computing*, vol. 28(12), páginas 1575–1589, 2010.

DOUCET, A., DE FREITAS, N. y GORDON, N. *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.

DOUCET, A., GODSILL, S. y ANDRIEU, C. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, vol. 10(3), páginas 197–208, 2000.

DREW, M., LI, Z. y ZHONG, X. Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences. En *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3, páginas 929–932. IEEE, 2002.

DUAN, L., JIN, J., TIAN, Q. y XU, C. Nonparametric motion characterization for robust classification of camera motion patterns. *Multimedia, IEEE Transactions on*, vol. 8(2), páginas 323–340, 2006.

DUMITRAS, A. y HASKELL, B. A look-ahead method for pan and zoom detection in video sequences using block-based motion vectors in polar coordinates. En *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on*, vol. 3. IEEE, 2004.

ELGAMMAL, A., DURAISWAMI, R., HARWOOD, D. y DAVIS, L. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, vol. 90(7), páginas 1151–1163, 2002.

ELGAMMAL, A., HARWOOD, D. y DAVIS, L. Non-parametric model for background subtraction. *Proceeding of the European Conference on Computer Vision 2000*, páginas 751–767, 2000.

ELHABIAN, S., EL-SAYED, K. y AHMED, S. Moving object detection in spatial domain using background removal techniques-state-of-art. *Recent Patents on Computer Science*, vol. 1(1), páginas 32–54, 2008.

DE LA ESCALERA, A. y ARMINGOL, J. Vehicle detection and tracking for visual understanding of road environments. *Robotica*, vol. 28(06), páginas 847–860, 2010.

FERNANDO, W., CANAGARAJAH, C. y BULL, D. Wipe scene change detection in video sequences. En *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, vol. 3, páginas 294–298. IEEE, 2002.

FOIX, S., ALENYA, G. y TORRAS, C. Lock-in time-of-flight (tof) cameras: A survey. *Sensors Journal, IEEE*, (99), páginas 1–1, 2011.

FU, X. y ZENG, J. An effective video shot boundary detection method based on the local color features of interest points. En *Electronic Commerce and Security, 2009. ISECS'09. Second International Symposium on*, vol. 2, páginas 25–28. IEEE, 2009.

GAO, H., THAM, J., XUE, P. y LIN, W. Complexity analysis of morphological area openings and closings with set union. *Image Processing, IET*, vol. 2(4), páginas 231–238, 2008.

GREGGIO, N., BERNARDINO, A. y SANTOS-VICTOR, J. Image Segmentation for Robots: Fast Self-adapting Gaussian Mixture Model. *Image Analysis and Recognition*, páginas 105–116, 2010.

GRIMSON, W. y STAUFFER, C. Adaptive background mixture models for real-time tracking. En *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 246–252. 1999.

HA, J. y LEE, W. Foreground objects detection using multiple difference images. *Optical Engineering*, vol. 49(4), página 7201, 2010.

HAN, B., COMANICIU, D., ZHU, Y. y DAVIS, L. Sequential kernel density approximation and its application to real-time visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, páginas 1186–1197, 2007.

HARRIS, C. y STEPHENS, M. A combined corner and edge detector. En *Alvey vision conference*, vol. 15, página 50. Manchester, UK, 1988.

HAUPTMANN, A., YAN, R., QI, Y., JIN, R., CHRISTEL, M., DERTHICK, M., CHEN, M., BARON, R., LIN, W. y NG, T. Video classification and retrieval with the informedia digital video library system. *NIST SPECIAL PUBLICATION SP*, páginas 119–127, 2003.

HORN, B. y SCHUNCK, B. Determining optical flow. *Artificial intelligence*, vol. 17(1-3), páginas 185–203, 1981.

HSIAO, Y., CHUANG, C., JIANG, J. y CHIEN, C. A contour based image segmentation algorithm using morphological edge detection. En *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 3, páginas 2962–2967. IEEE, 2006.

HUAN, Z., XIUHUAN, L. y LILEI, Y. Shot boundary detection based on mutual information and canny edge detector. En *2008 International Conference on Computer Science and Software Engineering*, páginas 1124–1128. IEEE, 2008.

IONESCU, B., BUZULOIU, V., LAMBERT, P. y COQUIN, D. Improved cut detection for the segmentation of animation movies. En *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 2. IEEE, 2006.

ISARD, M. y BLAKE, A. Condensation-conditional density propagation for visual tracking. *International Journal of Computer Vision*, vol. 29(1), páginas 5–28, 1998.

JEON, H., BASSO, A. y DRIESSEN, P. Camera motion detection in video sequences using motion cooccurrences. *Advances in Mulitmedia Information Processing-PCM 2005*, páginas 524–534, 2005.

JIAYIN, L., CHUANG, W. y KIM, J. Camera motion detection for conversation scenes in movies. En *Computational and Information Sciences (ICCIS), 2010 International Conference on*, páginas 725–728. IEEE, 2010.

JOYCE, R. y LIU, B. Temporal segmentation of video using frame and histogram space. *Multimedia, IEEE Transactions on*, vol. 8(1), páginas 130–140, 2006.

KEKRE, H., SARODE, T. y RAUL, B. Color image segmentation using Kekre's fast codebook generation algorithm based on energy ordering concept. En *Proceedings of the International Conference on Advances in Computing, Communication and Control*, páginas 357–362. ACM, 2009.

KELKAR, D. y GUPTA, S. Improved quadtree method for split merge image segmentation. En *First International Conference on Emerging Trends in Engineering and Technology*, páginas 44–47. IEEE, 2008.

KHAN, Z., BALCH, T. y DELLAERT, F. An MCMC-based particle filter for tracking multiple interacting targets. *Computer Vision-ECCV 2004*, páginas 279–290, 2004.

KOPRINSKA, I. y CARRATO, S. Temporal video segmentation: A survey. *Signal processing: Image communication*, vol. 16(5), páginas 477–500, 2001.

KUCUKTUNC, O., GUDUKBAY, U. y ULUSOY, O. Fuzzy color histogram-based video segmentation. *Computer Vision and Image Understanding*, vol. 114(1), páginas 125–134, 2010.

KWON, C., HAN, D., KIM, H., LEE, M. y PARK, S. Dissolve Detection Using Intensity Change Information of Edge Pixels. *IEICE Transactions on Information and Systems*, vol. 91(1), página 153, 2008.

LANDABASO, J. y PARDAS, M. A unified framework for consistent 2-d/3-d foreground object detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18(8), páginas 1040–1051, 2008.

LAWRENCE, S., ZIOU, D., AUCLAIR-FORTIER, M., WANG, S. ET AL. Motion-Insensitive Detection of Cuts and Gradual Transitions in Digital Video. *Pattern Recognition and Image Analysis*, vol. 14(1), páginas 109–119, 2004.

LEE, H., LEE, C. y KIM, S. Abrupt shot change detection using an unsupervised clustering of multiple features. En *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 6, páginas 2015–2018. IEEE, 2000.

LEE, M., NEPAL, S. y SRINIVASAN, U. Edge-based semantic classification of sports video sequences. En *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1. IEEE, 2003.

LEE, S., KIM, Y. y CHOI, S. Fast scene change detection using direct feature extraction from MPEG compressed videos. *Multimedia, IEEE Transactions on*, vol. 2(4), páginas 240–254, 2002.

LEE, W. y LEE, M. A multi-class object classifier using boosted gaussian mixture model. En *Neural Information Processing. Theory and Algorithms*, vol. 6443 de *Lecture Notes in Computer Science*, páginas 430–437. Springer Berlin / Heidelberg, 2010.

LEFÈVRE, S., HOLLER, J. y VINCENT, N. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, vol. 9(1), páginas 73–98, 2003.

LEW, M., SEBE, N., DJERABA, C. y JAIN, R. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 2(1), página 19, 2006.

LI, J., DING, Y., SHI, Y. y LI, W. A divide-and-rule scheme for shot boundary detection based on sift. *Journal of JDCTA, International Journal of Digital Content Technology and its Applications*, vol. 4(3), páginas 202–214, 2010.

LIAO, J., WU, Y. y LIN, Y. Improving Sheather and Jones'bandwidth selector for difficult densities in kernel density estimation. *Journal of Nonparametric Statistics*, vol. 22(1), páginas 105–114, 2010.

LIN, G., CHANG, M. y CHIU, S. Dissolve Detection Scheme with Transition Duration Refinement. En *Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on*, vol. 1, páginas 155–158. IEEE, 2008.

LIN, S. y LEE, S. Using chromaticity distributions and eigenspace analysis for pose-, illumination-, and specularity-invariant recognition of 3d objects. En *Computer Vision and Pattern Recognition, 1997 IEEE Computer Society Conference on*, páginas 426–431. IEEE, 1997.

LIN, W., SUN, M., LI, H. y HU, H. A new shot change detection method using information from motion estimation. *Advances in Multimedia Information Processing-PCM 2010*, páginas 264–275, 2011.

LIU, T., ZHANG, X., WANG, D., FENG, J. y LO, K. Inertia-based cut detection technique: a step to the integration of video coding and content-based retrieval. En *Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on*, vol. 2, páginas 1018–1025. IEEE, 2002.

LO, B. y VELASTIN, S. Automatic congestion detection system for underground platforms. En *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, páginas 158–161. IEEE, 2002.

LU, L. y HAGER, G. A nonparametric treatment for location/segmentation based visual tracking. En *IEEE Conference on Computer Vision and Pattern Recognition*, páginas 1–8. 2007.

MAITRA, M. y CHATTERJEE, A. A hybrid cooperative-comprehensive learning based PSO algorithm for image segmentation using multilevel thresholding. *Expert Systems with Applications*, vol. 34(2), páginas 1341–1350, 2008.

MARTEL-BRISSON, N. y ZACCARIN, A. Unsupervised approach for building non-parametric background and foreground models of scenes with significant foreground activity. En *Proceeding of the 1st ACM workshop on Vision networks for behavior analysis*, páginas 93–100. ACM, 2008.

MEIER, E. y ADE, F. Using the condensation algorithm to implement tracking for mobile robots. En *Advanced Mobile Robots, 1999.(Eurobot'99) 1999 Third European Workshop on*, páginas 73–80. IEEE, 1999.

MICHELONI, C., RINNER, B. y FORESTI, G. Video analysis in pan-tilt-zoom camera networks. *Signal Processing Magazine, IEEE*, vol. 27(5), páginas 78–90, 2010.

MITTAL, A. y PARAGIOS, N. Motion-based background subtraction using adaptive kernel density estimation. En *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004.

MOHANTY, S. Gpu-cpu multi-core for real-time signal processing. En *Consumer Electronics, 2009. ICCE'09. Digest of Technical Papers International Conference on*, páginas 1–2. IEEE, 2009.

NAGASAKA, A. y TANAKA, Y. Automatic video indexing and full-video search for object appearances (abstract). *Journal of Information Processing*, vol. 15(2), página 316, 1992.

NASCIMENTO, J., FIGUEIREDO, M. y MARQUES, J. Trajectory classification using switched dynamical hidden Markov models. *Image Processing, IEEE Transactions on*, vol. 19(5), páginas 1338–1348, 2010.

NIETO, M., CUEVAS, C. y SALGADO, L. Measurement-based reclustering for multiple object tracking with particle filters. En *Image Processing (ICIP), 2009 16th IEEE International Conference on*, páginas 4097–4100. IEEE, 2010.

NIETO, M., CUEVAS, C., SALGADO, L. y GARCÍA, N. Line segment detection using weighted mean shift procedures on a 2D slice sampling strategy. *Pattern Analysis & Applications*, vol. 14, páginas 1–15, 2011.

PADALKAR, M. y ZAVERI, M. Dissolve detection based shot identification using singular value decomposition. En *2010 Fourth Asia International Conference on Mathematical/-Analytical Modelling and Computer Simulation*, páginas 312–316. IEEE, 2010.

PAPOULIS, A. y PILLAI, S. *Probability, random variables, and stochastic processes*. McGraw-Hill, 2002.

PARKS, D. y FELS, S. Evaluation of background subtraction algorithms with post-processing. En *Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on*, páginas 192–199. IEEE, 2008.

PARZEN, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, vol. 33(3), páginas 1065–1076, 1962.

PENG, W., CHU, W., CHANG, C., CHOU, C., HUANG, W., CHANG, W. y HUNG, Y. Editing by viewing: Automatic home video summarization by viewing behavior analysis. *Multimedia, IEEE Transactions on*, vol. 13(3), páginas 539–550, 2011.

PÉREZ, A. y GONZÁLEZ, R. An iterative thresholding algorithm for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), páginas 742–751, 2009.

PICCARDI, M. Background subtraction techniques: a review. En *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 4, páginas 3099–3104. 2005.

PLOTKOWIAK, T. y LAY, J. Multi-feature Multi-pass Dissolve Detection. En *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on*, páginas 332–339. IEEE, 2008.

PROAKIS, J. G. *Ditital Communications*. McGraw-Hill, 2001.

PYE, D., HOLLINGHURST, N., MILLS, T. y WOOD, K. Audio-visual segmentation for content-based retrieval. En *Fifth International Conference on Spoken Language Processing*. Citeseer, 1998.

RAMOS, V. y MUGE, F. Map segmentation by colour cube genetic K-mean clustering. *Research and Advanced Technology for Digital Libraries*, páginas 319–323, 2010.

REN, J., JIANG, J., CHEN, J. y IPSON, S. Extracting objects and events from mpeg videos for highlight-based indexing and retrieval. *Journal of Multimedia*, vol. 5(2), páginas 95–103, 2010.

REN, W., SHARMA, M. y SINGH, S. Automated video segmentation. En *International Conference on Information, Communication, and Signal Processing, Singapore*. Citeseer, 2001.

REYNOLDS, M., DOBOŠ, J., PEEL, L., WEYRICH, T. y BROSTOW, G. Capturing time-of-flight data with confidence. En *Proc. CVPR*, vol. 1377, páginas 945–952. 2011.

ROSENBLATT, M. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, vol. 27(3), páginas 832–837, 1956.

ROSIN, P. Thresholding for change detection. En *Computer Vision, 1998. Sixth International Conference on*, páginas 274–279. IEEE, 1998.

ROYDEN, C. y CONNORS, E. The detection of moving objects by moving observers. *Vision Research*, vol. 50(11), páginas 1014–1024, 2010.

SAHA, S. y MAULIK, U. A new line symmetry distance based automatic clustering technique: Application to image segmentation. *International Journal of Imaging Systems and Technology*, vol. 21(1), páginas 86–100, 2011.

SALEMBIER, P. y WILKINSON, M. Connected operators. *Signal Processing Magazine, IEEE*, vol. 26(6), páginas 136 –157, 2009.

SÁNCHEZ, J. y BINEFA, X. Shot segmentation using a˜ coupled markov chains representation of video contents. *Pattern Recognition and Image Analysis*, páginas 902–909, 2003.

SANGANI, K. 2010 gadget census [consumer tech census]. *Engineering & Technology*, vol. 5(14), páginas 28–29, 2010.

SATHYA, P. y KAYALVIZHI, R. Modified bacterial foraging algorithm based multilevel thresholding for image segmentation. *Engineering Applications of Artificial Intelligence*, vol. 24(4), páginas 595–615, 2011.

SEIDL, M., ZEPPELZAUER, M. y BREITENEDER, C. A study of gradual transition detection in historic film material. En *Proceedings of the second workshop on eHeritage and digital art preservation*, páginas 13–18. ACM, 2010.

SENTHILKUMARAN, N. y RAJESH, R. Edge Detection Techniques for Image Segmentation-A Survey of Soft Computing Approaches. *International Journal of Recent Trends in Engineering*, vol. 1(2), páginas 250–254, 2009.

SERBY, D., KOLLER-MEIER, E. y VAN GOOL, L. Probabilistic object tracking using multiple features. *Pattern Recognition*, vol. 2, páginas 184–187, 2004.

SHAH, S., NASEEM, U. y KARIM, A. Motion Segmentation through Incremental Hierarchical Clustering. En *Multitopic Conference, 2006. INMIC'06. IEEE*, páginas 134–139. IEEE, 2007.

SHAPIRO, L. y STOCKMAN, G. *Computer Vision*. Prentice Hall Upper Saddle River, New York, 2001.

SHEIKH, Y., JAVED, O. y KANADE, T. Background subtraction for freely moving cameras. En *Computer Vision, 2009 IEEE 12th International Conference on*, páginas 1219–1225. IEEE, 2009.

SHEIKH, Y. y SHAH, M. Bayesian modeling of dynamic scenes for object detection. *IEEE transactions on pattern analysis and machine intelligence*, vol. 27(11), páginas 1778–1792, 2005.

SMAL, I., NIESSEN, W. y MEIJERING, E. Advanced particle filtering for multiple object tracking in dynamic fluorescence microscopy images. En *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, páginas 1048–1051. IEEE, 2007.

SMEATON, A., OVER, P. y DOHERTY, A. Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, vol. 114(4), páginas 411–418, 2010.

SNOEK, C. y WORRING, M. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, vol. 25(1), páginas 5–35, 2005.

STAUFFER, C. y GRIMSON, W. Adaptive background mixture models for real-time tracking. En *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2. 2002a.

STAUFFER, C. y GRIMSON, W. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22(8), páginas 747–757, 2002b.

SUHR, J., JUNG, H., NOH, S. y KIM, J. Background compensation for pan-tilt-zoom cameras using 1-d feature matching and outlier rejection. *Circuits and Systems for Video Technology, IEEE Transactions on*, (99), páginas 1–1, 2011.

TAN, I., VAN SCHIJNDEL, R., FAZEKAS, F., FILIPPI, M., FREITAG, P., MILLER, D., YOUSRY, T., POUWELS, P., ADÈR, H. y BARKHOF, F. Image registration and subtraction to detect active t2 lesions in ms: an interobserver study. *Journal of neurology*, vol. 249(6), páginas 767–773, 2002.

TANAKA, T., SHIMADA, A., TANIGUCHI, R., YAMASHITA, T. y ARITA, D. Towards robust object detection: integrated background modeling based on spatio-temporal features. *Computer Vision–ACCV 2009*, páginas 201–212, 2010.

TANG, Z., MIAO, Z. y WAN, Y. Background Subtraction Using Running Gaussian Average and Frame Difference. *Entertainment Computing–ICEC 2007*, páginas 411–414, 2007.

TAO, K., LIN, S. y ZHANG, Y. Compressed domain motion analysis for video semantic events detection. En *2009 WASE International Conference on Information Engineering*, páginas 201–204. IEEE, 2009.

TAUB, H. y SCHILLING, D. *Principles of communication systems*. McGraw-Hill Higher Education, 1986.

TAVAKKOLI, A., NICOLESCU, M., BEBIS, G. y NICOLESCU, M. Non-parametric statistical background modeling for efficient foreground region detection. *Machine Vision and Applications*, vol. 20(6), páginas 395–409, 2009.

TENG, S., TAN, W. y HUANG, G. Video temporal segmentation using cooperative model. En *Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on*, páginas 201–206. IEEE, 2008.

TONOMURA, Y. y ABE, S. Content oriented visual interface using video icons for visual database systems. *Journal of Visual Languages & Computing*, vol. 1(2), páginas 183–198, 1990.

TOYAMA, K., KRUMM, J., BRUMITT, B. y MEYERS, B. Wallflower: principles and practice of background maintenance. En *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, páginas 255 –261 vol.1. Disponible en `http://research.microsoft.com/en-us/um/people/jckrumm/wallflower/testimages.htm/`.

TRAN, S., LIN, Z., HARWOOD, D. y DAVIS, L. Umd_vdt, an integration of detection and tracking methods for multiple human tracking. *Multimodal Technologies for Perception of Humans*, páginas 179–190, 2009.

TRUONG, B. y VENKATESH, S. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 3(1), página 3, 2007.

TSAI, S., CHENG, C., LI, C. y CHEN, L. A real-time 1080p 2d-to-3d video conversion system. En *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, páginas 803–804. IEEE, 2011.

TSENG, V., CHEN, C., CHEN, C. y HONG, T. Segmentation of time series by the clustering and genetic algorithms. En *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, páginas 443–447. IEEE, 2006.

TURLACH, B. Bandwidth selection in kernel density estimation: A review. *CORE and Institut de Statistique*, páginas 23–493, 1993.

UTASI, A. y CZUNI, L. HMM-based unusual motion detection without tracking. En *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, páginas 1–4. IEEE, 2009.

VARSHNEY, S., RAJPAL, N. y PURWAR, R. Comparative study of image segmentation techniques and object matching using segmentation. En *Methods and Models in Computer Science, 2009. ICM2CS 2009. Proceeding of International Conference on*, páginas 1–6. IEEE, 2010.

WAN, Q. y WANG, Y. Background subtraction based on adaptive non-parametric model. En *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, páginas 5960–5965. IEEE, 2008.

WAND, M. y JONES, M. Kernel smoothing, volume 60 of Monographs on Statistics and Applied Probability. *Chapman Hall, New York*, 1995.

WANG, C., ZHANG, X., YUAN, C. y LIU, Y. Video Segmentation of Illuminance Abrupt Variation Based on MOGs and Interframe Gradient Cross-correlation. En *Signal Processing, 2006 8th International Conference on*, vol. 1. IEEE, 2007.

WANG, H. y OLIENSIS, J. Generalizing edge detection to contour detection for image segmentation. *Computer Vision and Image Understanding*, vol. 114(7), páginas 731–744, 2010.

WANG, H. y SUTER, D. A re-evaluation of mixture of Gaussian background modeling [video signal processing applications]. En *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2. IEEE, 2005.

WANG, L., HU, W. y TAN, T. Recent developments in human motion analysis. *Pattern recognition*, vol. 36(3), páginas 585–601, 2003.

WANG, L., WU, H. y PAN, C. Adaptive εLBP for background subtraction. *Computer Vision–ACCV 2010*, páginas 560–571, 2011.

WANG, Y., YANG, Y., REN, T. y WU, G. A motion-insensitive dissolve detection method with surf. En *2009 Fifth International Conference on Image and Graphics*, páginas 451–456. Ieee, 2009.

WARHADE, K., MERCHANT, S. y DESAI, U. Effective algorithm for detecting various wipe patterns and discriminating wipe from object and camera motion. *Image Processing, IET*, vol. 4(6), páginas 429–442, 2010.

WHITE, B. y SHAH, M. Automatically tuning background subtraction parameters using particle swarm optimization. En *Multimedia and Expo, 2007 IEEE International Conference on*, páginas 1826–1829. IEEE, 2007.

YAO, W., ZHAI, G. y CAI, J. An effective dissolve detection approach with temporal and spatial considerations. En *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*, páginas 1283–1286. IEEE, 2008.

YU, Q. y CLAUSI, D. SAR sea-ice image analysis based on iterative region growing using semantics. *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45(12), páginas 3919–3931, 2007.

YU, Q. y CLAUSI, D. Irgs: Image segmentation using edge penalties and region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30(12), páginas 2126–2139, 2008.

YUAN, X. y FENG, H. An abrupt shot change detection algorithm based on the yuv space. En *2010 International Conference on Electrical and Control Engineering*, páginas 4630–4633. IEEE, 2010.

YUFENG, L., YINGHUA, Y. y GUIJU, L. A novel wipe transition detection method based on multi-feature. En *2010 Third International Conference on Knowledge Discovery and Data Mining*, páginas 451–454. IEEE, 2010.

ZANG, Q. y KLETTE, R. Parameter analysis for mixture of gaussians model. Informe técnico, CITR, The University of Auckland, New Zealand, 2006.

ZHANG, H., KANKANHALLI, A. y SMOLIAR, S. Automatic partitioning of full-motion video. *Multimedia Systems*, vol. 1(1), páginas 10–28, 1993.

ZHANG, H., LOW, C. y SMOLIAR, S. Video parsing and browsing using compressed data. *Multimedia tools and applications*, vol. 1(1), páginas 89–111, 1995.

ZHANG, X. y YANG, J. Foreground segmentation based on selective foreground model. *Electronics Letters*, vol. 44(14), páginas 851–852, 2008.

ZHOU, D. y ZHANG, H. Modified GMM background modeling and optical flow for detection of moving objects. En *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 3, páginas 2224–2229. IEEE, 2006.

–¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.

Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes